Psychometrika

VOLUME XVII—1952
JANUARY-DECEMBER

PUBLISHED QUARTERLY
By THE PSYCHOMETRIC SOCIETY
AT 1407 SHERWOOD AVENUE
RICHMOND 5, VIRGINIA

Faxon

Psychometrika

CONTENTS

VOLUME SEVENTEEN MARCH 1952 NUMBER ONE

BIOSTATISTICS CONFERENCE
June 16-July 23, 1952

Iowa State College, Ames, Iowa

A biostatistics conference has been scheduled for the first session of the 1952 Summer Quarter at Ames, Iowa, sponsored by faculty members working in agriculture, biology, and statistics at Iowa State College and by The Biometric Society (ENAR). The five-week conference is arranged so that persons who cannot attend the entire conference can advantageously come for one or more weeks. Iowa State College is giving the Conference financial support.

The plan is that each morning a biologist will present a problem, outline the objectives, and describe techniques suitable for the experiment and analysis. A paired statistician will discuss suitable experimental designs and statistical and mathematical methods for attacking the problem. These speakers will preside at a general discussion period of the same topic the same afternoon.

The program is tentatively arranged in five somewhat separate weekly units as follows:

First week: **Development of Quantitative Biology**

Second week: **Specification of Populations and Their Processes**

Third week: **The Estimation of Populations**

Fourth week: **Individual Growth**

Fifth week: Biomathematical Mechanisms Within the Individual and Species

It is expected that the Conference will be of interest to advanced undergraduates, graduate students, and research workers in the various biological sciences and to statisticians who are interested in statistics as a research tool. Some graduate credit in Statistics at Iowa State College will be allowed for attendance and study during the Conference.

Rooms will be available in the college dormitories at the usual rates. For more detailed information write: T. A. Bancroft, Director, Statistical Laboratory, Iowa State College, Ames, Iowa.

PSYCHOMETRIKA—VOL. 17, NO. 1
MARCH, 1952

THE SELECTIVE EFFICIENCY OF A TEST BATTERY*

HERBERT S. SICHEL

NATIONAL INSTITUTE FOR PERSONNEL RESEARCH, SOUTH AFRICAN COUNCIL FOR SCIENTIFIC AND INDUSTRIAL RESEARCH

In industrial acceptance sampling one frequently makes use of operating characteristic curves to describe the discriminating power of a particular sampling plan. Similarly, it is possible to demonstrate the selective efficiency of a test battery in terms of (a) the Applicant's Operating Characteristic (A.O.C.); (b) the Selector's Operating Characteristic (S.O.C). The A.O.C. determines the chance of selection by means of a test for any given level of true ability. The S.O.C. connects functionally probability of success on the criterion with the predictor scores of a battery. For the case of a normal bivariate distribution the exact mathematical expressions of the OC curves are derived in terms of the correlation coefficient ρ, the cut-off points α and β, and the predictor and criterion scores X and Y (in standard measures). The Efficiency Index H is defined as the percentage of successful subjects gained by the use of a test battery, taking chance selection as a yardstick for comparison. Its optimum, for fixed ρ and α, is derived. The distribution law of the criterion scores of selectees is deduced and its first four moments are shown to depart little from normality for cases usually encountered in practice. A "Quality-Gain" diagram graphically illustrates the improvements secured. Another simple device, the "Cost-Utility" diagram, explains to management the full implications of selecting personnel by means of a test battery. Neither of the diagrams requires an understanding of the correlation coefficient. The confidence belt of the OC curves, the standard error of the mean criterion score of selectees and the standard error of the predicted number of successful applicants are determined. Finally, the full theory is applied in detail to a real test battery.

In a large scale selection program we may distinguish three parties all having somewhat different approaches and interests at heart. They are:

(a) The men seeking employment, entrance to a college, scholarships, induction into an army, etc. We shall call them the *applicants*.

(b) The selection agency, which may be a personnel department, a draft board, a selection committee, or a psychological unit hereafter called the *selector*.

(c) The agency for which selection is to be carried out such as management, the army, a university, in short, the *employer*.

*This paper is the result of an investigation undertaken by the South African National Institute for Personnel Research, Johannesburg, and was completed at the Educational Testing Service, Princeton. The author wishes to thank these organizations for permission to publish this paper.

In the following, an attempt has been made to describe quantitatively what risks the various parties run in going through or having installed a selection program. The principal instruments of measurement are the applicant's and selector's operating characteristics (OC curves), the efficiency index of a selection procedure, the quality-gain diagram, and the cost-utility diagram.

I. *The Mathematical Model*

(*a*) *The Applicant's Operating Characterictic* (*A.O.C. Curve*)

Whether it is a job in an industrial enterprise or admission to a pilot training course, the individual applicant's main interest lies in being picked to the vacancy for which he is competing with other candidates. When turned down, he may question the competence of the selector, feeling that his true potential was not recognized. Especially aptitude-test selection has come in for a lot of fire. It is said that this procedure is wholly undemocratic as it takes no account of the individual. It is supposed to work only in the employer's interest as it creams off the best test scorers but does not give any consideration to the individual who may have all abilities requisite for the job but is unlucky to be below the cut-off point on the battery.

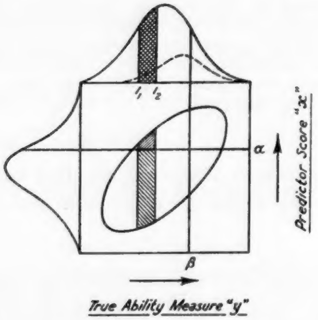

FIG. 1 BIVARIATE DISTRIBUTION

Let us assume that predictor scores and true criterion ability measures are normally correlated as shown in Figure 1. For applicants whose true ability is

$$l_1 \leq y \leq l_2 ,$$

the probability of scoring equal to or above the predictor cut-off point

a is clearly

$$\text{prob.}\,(x \geqslant a) = \frac{\int\limits_{l_1}^{l_2}\int\limits_{a}^{+\infty} f(x,y)\,dy\,dx}{\int\limits_{l_1}^{l_2}\int\limits_{-\infty}^{+\infty} f(x,y)\,dy\,dx}, \tag{1}$$

where

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)}$$

and both variables expressed as standard measures. Equation (1) may be simplified by substitution of

$$u = \frac{x-\rho y}{\sqrt{1-\rho^2}}.$$

It becomes

$$\text{prob.}\,(x \geq a) = \frac{1}{\sqrt{2\pi}}\,\frac{\int\limits_{l_1}^{l_2} e^{-\frac{y^2}{2}} \left| \int\limits_{\frac{a-\rho y}{\sqrt{1-\rho^2}}}^{+\infty} e^{-\frac{u^2}{2}}\,\mathrm{du} \right| dy}{\int\limits_{l_1}^{l_2} e^{-\frac{y^2}{2}}\,dy}. \tag{2}$$

Suppose now that the interval $l_2 - l_1$ is small but finite. Then, in view of the mean value theorem,

$$\text{prob.}\,(x \geq a) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(\theta_1^2-\theta_2^2)} \int\limits_{-\frac{a-\rho\theta_1}{\sqrt{1-\rho^2}}}^{+\infty} e^{-\frac{u^2}{2}}\,du\,, \tag{3}$$

where

$$l_1 < \theta_1 < l_2\,,$$

$$l_1 < \theta_2 < l_2\,.$$

Keeping l_2 constant and letting l_1 approach l_2 we have

$$\theta_1 \rightarrow l_2\,,$$

$$\theta_2 \rightarrow l_2\,,$$

and in the limit (3) becomes

$$\lim_{l_1 \rightarrow l_2} .\ \text{prob.}\,(x \geq a) = \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\rho l_2}{\sqrt{1-\rho^2}}}^{+\infty} \mathrm{e}^{-\frac{u^2}{2}}\, du\,. \tag{4}$$

Replacing l_2 in equation (4) by y, we find for the conditional probability

$$\text{prob.}\,(x \geq a|y) = \pi(y) = \Phi\left(\frac{a-\rho y}{\sqrt{1-\rho^2}}\right), \tag{5}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{+\infty} e^{-\frac{u^2}{2}}\, du\,, \text{ the standard normal integral.}$$

Equation (5) is called the Applicant's Operating Characteristic giving an individual's chance of selection to the vacancy* as a function of his true ability measure.

Equation (5) could have been derived in a simpler fashion. However the limit approach is to be preferred as it produces formula (2) as a valuable by-product, i.e., the chance of selection of a group of applicants all of whom fall within an ability interval.

If $\rho = 0$, that is if we deal with a pure chance selection, (5) degenerates into

$$\pi(y) = \Phi(a) = \text{constant}\,.$$

The A.O.C. curves for $a = 1.0$, $\rho = .6$ and $\rho = 0$ have been drawn in Figure 2. On the abscissa standard scores ($\xi = 50$, $\sigma = 10$) were plotted instead of standard measures. ("Standard Score" is symbolized in the diagrams by "*S/S.*") The graph clearly shows that an individual's probability of being selected is steadily increasing with increasing true ability. The higher a man's true potentialities to cope

*This follows directly from the fact that an applicant will be selected if his score $x \geq \alpha$. α, of course, is dependent on the selection ratio.

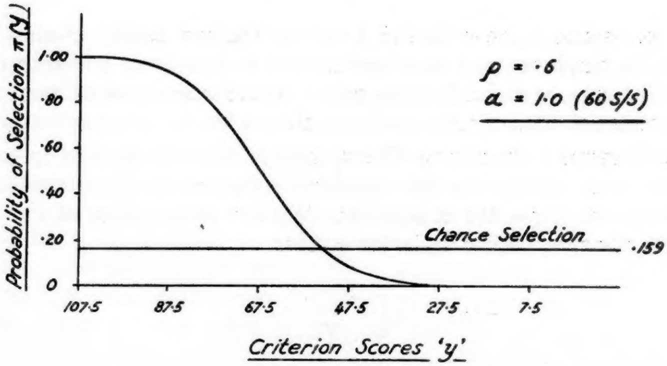

FIG. 2 APPLICANT'S OPERATING CHARACTERISTIC

with the job the better are the odds of his being chosen. On the other hand, chance selection does not take into account true abilities. As may be seen from Figure 2, the chance of selection is the same for the dull and the gifted.

It must not be inferred from the above that all non-test procedures of selection produce pure chance A.O.C. curves. However experimental evidence goes to show that non-test techniques give rise to very shallow A.O.C. curves.

For individuals of low true ability the chances of getting the job through test selection are minimal. Such candidates' rejection is in their own interest as it will save them disappointments and frustrations at a later stage when they turn out to be occupational misfits. We may, therefore, conclude that far from being undemocratic it is in the individual's own interest to be selected or rejected by means of a test battery.

(*b*) *The Selector's Operating Characteristic* (*S.O.C. Curve*)

After the construction of a reliable and valid test battery, the selector's main interest is focussed on the correctness of his selections. As correlations near unity are non-existent in the psychological prediction field, it will happen quite often that those recommended to the vacancies will not make the grade on the true ability variable. What we need is the assessment of risk we run in selecting an applicant with a given test or battery score x . The individual applicant's chance of success on the true ability scale is the probability of equalling or exceeding a lower bound β of satisfactory ability. β may be fixed by the employer at a standard known to him from previous experience; it may be defined as the mean criterion score of an untruncated sample; or it may arbitrarily be chosen by the selector himself.

The Selector's Operating Characteristic may be derived in precisely the same manner as the Applicant's Operating Characteristic. Reversal of the variables in equation (5) and substitution of cut-off point β on the true ability scale for α gives

$$\pi(x) = \Phi\left(\frac{\beta - \rho x}{\sqrt{1-\rho^2}}\right) , \tag{6}$$

which expresses the chance of success on the criterion variable as a function of test scores. In Figure 3 the S.O.C. curve has been drawn

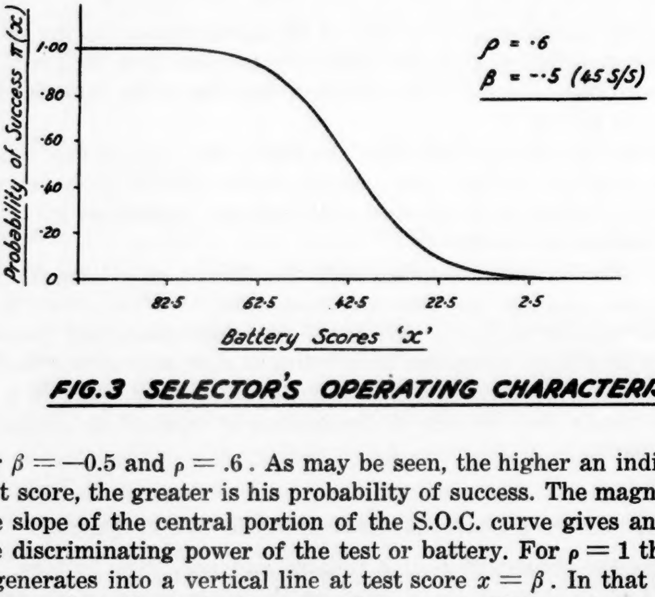

FIG. 3 SELECTOR'S OPERATING CHARACTERISTIC

for $\beta = -0.5$ and $\rho = .6$. As may be seen, the higher an individual's test score, the greater is his probability of success. The magnitude of the slope of the central portion of the S.O.C. curve gives an idea of the discriminating power of the test or battery. For $\rho = 1$ the curve degenerates into a vertical line at test score $x = \beta$. In that case we have perfect discrimination provided $\alpha = \beta$. If the cut-off point on the true ability scale (or its estimate, the criterion score) is changed to β', we shift the S.O.C. curve by $\dfrac{\beta' - \beta}{\rho}$ without altering its shape.

The S.O.C. curve based on equation (6) supplies the mathematical model for expectancy charts. It is useful for prediction per se, for graduating observed expectancies, and for testing statistically whether observed proportions of successes conform with or deviate from the hypotheses underlying expectancy charts. Equation (6) is not new. It has been used by McClelland (1), tables of it were computed by Bittner and Wilder (2), and the whole subject of expectancy charts has been well discussed by Bingham (3) recently.

In practice, observed proportions of successes are almost invariably based on grouped data. If the total range of test scores has been subdivided into ten or more intervals no serious error will be committed by substituting the midpoints or the exact mean values of the class intervals into equation (6) for making comparisons with the theoretical values. However, current expectancy charts are frequently based on five or fewer intervals. In that case it is safer to use the exact formula which we may write down from equation (2) by interchanging the variables and replacing α by β. It is

$$\text{prob.}(y \geq \beta) = \frac{1}{\sqrt{2\pi}} \frac{\int_{l_1}^{l_2} e^{-\frac{x^2}{2}} \left[\int_{\frac{\beta-\rho x}{\sqrt{1-\rho^2}}}^{+\infty} e^{-\frac{u^2}{2}} du \right] dx}{\int_{l_1}^{l_2} e^{-\frac{x^2}{2}} dx}. \qquad (7)$$

Numerical values of (7) may be found either from Pearson's *Tables for Statisticians and Biometricians* (Vol. II) or by a mechanical quadrature.

(*c*) *The Efficiency of a Test Battery*

(*1*) *The Efficiency Index*:The efficiency of a battery is the employer's chief concern. It is he who pays for the selection program and naturally, he is right to ask: Does it really pay?

Various devices have been proposed to answer his question. Taylor and Russell (4) suggested a comparison of the proportion of successful selectees with those obtained from batteries with zero and perfect validities. Brogden (5) introduced a measure of gain being the absolute difference of the criterion means of a selected and unselected population of applicants. A similar method, only on a percentage basis, had previously been described by Jarrett (6). Brogden (5) also considers the gain in income from improved production when the costs of large scale testing are taken into account. McClelland (1) uses the proportion of "misfits," this being the proportion of incorrectly accepted and incorrectly rejected applicants as an index of a battery's efficiency. He proves that for a given β the proportion of misfits is a minimum if

$$\Phi\left(\frac{\beta-\rho\,\alpha}{\sqrt{1-\rho^2}}\right)=\frac{1}{2}$$

and determines the predictor cut-off point α accordingly.

All the above measures serve useful purposes. However, in personnel selection the cost of testing is not always of importance, i.e., for scholarships or college entrance, and the concept of output gain needs considerable stretching when applied to pilot selection. The minimum number of misfits is an excellent yardstick in classification problems but becomes of dubious value in a selection program. Apparently the most direct measure of efficiency and the one which makes sense in the majority of practical cases is the one originally proposed by Taylor and Russell. It is developed further here.

A good device of efficiency would be the percentage gain in number of successful battery selected applicants over existing routine selection practice. Very rarely do we have data to measure the success (or otherwise) of non-test procedures. In an experimental setup it should be possible to have an untruncated sample followed up on which test and non-test selection has been carried out prior to induction into the work process. In such an experiment it should be possible to construct operating characteristics for both selection procedures.

Besides the lack of data, there is a further point to militate against the use of non-test procedures in gauging the efficacy of test selection. Whereas for a given applicant population and a stable criterion cut-off point β, battery selection will produce comparable results from place to place, this should not prove the case in a non-test procedure in which the success of selection depends entirely on the skill and human insight of the person hiring labour. For a scientific measure of efficiency we need a fixed bench mark with which we can compare the results of aptitude testing. It is for this reason that we have to fall back on pure chance selection, and we do not imply that non-test procedures do by necessity approach a lottery.

We shall define the "efficiency index" of a battery as the percentage gain of successful test selected applicants over chance selected candidates. Expressed in other words, it is the number of successful candidates gained over chance selection for each hundred applicants selected. We write

$$\text{Efficiency Index } H = 100\,(\pi_2 - \pi_1)\,\%\,, \tag{8}$$

where

$$\pi_2 = \frac{1}{\Phi(\alpha)} \int_{\beta}^{+\infty} \int_{\alpha}^{+\infty} f(x,y)\; dy\; dx \tag{8a}$$

and

$$\pi_1 = \Phi(\beta). \tag{8b}$$

$f(x,y)$ is again the normal bivariate function. Equation (8a) may be evaluated with Pearson's Tables or with a mechanical quadrature formula.

The predictor cutting score α will be dependent on the selection ratio and, as Taylor and Russell have previously shown, the lower it is (and consequently the larger α) the greater will be the efficiency index H.

The criterion cutting score β is usually fixed by the employer. However, where no such standard is known *a priori,* it is possible to locate β in such a way that H becomes a maximum. We have for maxima or minima

$$\frac{\partial H}{\partial \beta} = 100\left(\frac{\partial \pi_2}{\partial \beta} - \frac{\partial \pi_1}{\partial \beta} \right) = 0.$$

Equation (8) may be written

$$H = \frac{100}{\sqrt{2\pi}\,\Phi(\alpha)} \int_{\beta}^{+\infty} e^{-\frac{y^2}{2}}\; \Phi\left(\frac{\alpha - \rho\, y}{\sqrt{1-\rho^2}} \right) dy - 100\,\Phi(\beta)\,,$$

$$H = \frac{100}{\Phi(\alpha)} \int_{\beta}^{+\infty} \operatorname{erf}(y)\,\pi(y)\,dy - 100\,\Phi(\beta)\,, \tag{9}$$

where

$$\operatorname{erf}(y) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}}\,.$$

Hence

$$\frac{\partial H}{\partial \beta} = 100\left[-\frac{1}{\Phi(a)} \operatorname{erf}(\beta)\pi(\beta) + \operatorname{erf}(\beta) \right] = 0,$$

$$\pi(\beta) = \Phi\left(\frac{a - \rho\beta}{\sqrt{1-\rho^2}} \right) = \Phi(a),$$

$$\frac{a - \rho\beta}{\sqrt{1-\rho^2}} = a$$

and finally for maximum condition

$$\beta_{\max} = \left(\frac{1 - \sqrt{1-\rho^2}}{\rho} \right) a. \tag{10}$$

The formal proof that we really deal with a maximum has been omitted. In Figure 4 equation (9) has been graphed for a = 1.0 and ρ = .6 . It brings out the important point that the efficiency of a battery is not only conditioned by the selection ratio and correlation coefficient but is also very much dependent on the criterion cutting

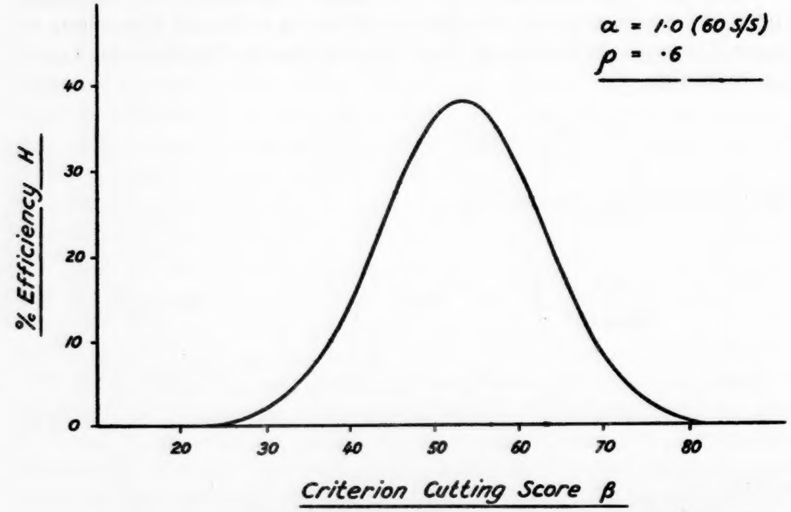

FIG. 4 BATTERY EFFICIENCY H AS A FUNCTION OF THE CRITERION CUTTING SCORE β

score β. For a given correlation and selection ratio the efficiency decreases rapidly on either side of the optimal cut-off point $\beta_{\max}$.

(2) *The Quality-Gain Diagram*: A comparison of the criterion distribution of the selected group of applicants (shown as the dotted curve in Figure 1) with that of an unselected group is of importance as it may be used to combine the usefulness of the efficiency index H with a graphical description of the over-all improvement in quality performance of battery selected personnel. The true ability distribution of the unselected group is

$$\operatorname{erf}(y)\,dy\,,$$

and the probability of being selected through a test is

$$\pi(y)=\Phi\left(\frac{a-\rho y}{\sqrt{1-\rho^2}}\right)$$

for a given level of true ability y. The total number of persons selected is $N\Phi(a)$. Hence the required probability distribution of criterion scores (being the estimate of true ability) is

$$\phi(y)\,dy=\frac{1}{\Phi(a)}\operatorname{erf}(y)\,\pi(y)\,dy. \tag{11}$$

Jarrett (6) has previously shown that the standard deviation of the criterion distribution of the selectees differs but little from the unrestricted criterion distribution. What is even more astonishing is the fact that the criterion distribution is staying nearly normal in spite of heavy truncation on the predictor variable. We may prove this phenomenon by deriving the first four moments of (11).

The mean criterion score of applicants having all predictor score x is

$$\bar{y}=\rho x\,,$$

and the mean criterion score of all selectees

$$M_1'=\frac{\rho}{\Phi(a)}\int_a^\infty x\operatorname{erf}(x)\,dx=\rho\, m_1'\,, \tag{12}$$

where m_1' is the mean of the truncated tail about the origin. The first four moments of individual array distributions about array means are

$$\left.\begin{aligned}\mu_1&=0\,,\\ \mu_2&=1-\rho^2,\\ \mu_3&=0\,,\\ \mu_4&=3(1-\rho^2)^2.\end{aligned}\right\} \tag{13}$$

We now take the moments of individual array distributions about M_1'. Central moments may be transferred to other axes by the well known formula

$$\nu_n' = \mu_n + nd\mu_{n-1} + \frac{n(n-1)}{2!} d^2\mu_{n-2} + \cdots,$$

where d is the distance between the centroid and the new axis. In our case

$$d = \bar{y} - M_1' = \rho(x - m_1'),$$

hence

$$\begin{aligned} \nu_n' = \mu_n &+ n\mu_{n-1}\rho(x - m_1') \\ &+ \frac{n(n-1)}{2!}\mu_{n-2}\rho^2(x - m_1')^2 + \cdots. \end{aligned} \tag{14}$$

Summing the arrays $x \geq a$, we find

$$\begin{aligned} M_n = \frac{1}{\Phi(a)} \int_a^{\infty} &\left[\mu_n + n\mu_{n-1}\rho(x - m_1') \right. \\ &\left. + \frac{n(n-1)}{2!}\mu_{n-2}\rho^2(x - m_1')^2 + \cdots \right] \\ &\times \operatorname{erf}(x)\,dx = \mu_n + \frac{n(n-1)}{2!}\mu_{n-2}\rho^2 m_2 \\ &+ \frac{n(n-1)(n-2)}{3!}\mu_{n-3}\rho^3 m_3 + \cdots, \end{aligned} \tag{15}$$

where m_n is the nth central moment of the predictor scores of the selected group. Substitution of (13) into (15) gives

$$\left. \begin{aligned} M_2 &= 1 - \rho^2 + \rho^2 m_2, \\ M_3 &= \rho^3 m_3, \\ M_4 &= 3(1-\rho^2)^2 + 6(1-\rho^2)\rho^2 m_2 + \rho^4 m_4, \end{aligned} \right\} \tag{16}$$

and finally for skewness and kurtosis of the $\phi(y)$ distribution,

$${}_y\beta_1 = \frac{\rho^6\, {}_x\beta_1}{\left(\dfrac{1-\rho^2}{m_2} + \rho^2 \right)^3}, \tag{17}$$

$${}_y\beta_2 = 3 + \frac{\rho^4({}_x\beta_2 - 3)}{\left(\dfrac{1-\rho^2}{m_2} + \rho^2 \right)^2}, \tag{18}$$

where subscripts x refer to the truncated predictor distribution, subscripts y to the resulting criterion distribution of selectees and

$$ {}_y\beta_1 = \frac{M_3^2}{M_2^3}, \qquad {}_x\beta_1 = \frac{m_3^2}{m_2^3}, $$

$$ {}_y\beta_2 = \frac{M_4}{M_2^2}, \qquad {}_x\beta_2 = \frac{m_4}{m_2^2}. $$

For the special case of $a = 0$,

$$ \left. \begin{aligned} m_2 &= \frac{\pi - 2}{\pi}, \\ {}_x\beta_1 &= \frac{2(4-\pi)^2}{(\pi-2)^3}, \\ {}_x\beta_2 &= \frac{3\pi^2 - 4\pi - 12}{(\pi-2)^2}. \end{aligned} \right\} \qquad (19) $$

Equations (16), (17), and (18) express the parameters of the criterion distribution of selectees as functions of similar parameters of the truncated predictor distribution. We may, therefore, compute numerical values by making use of Pearson's Tables of the moments of a truncated normal distribution. Table 1 gives some numerical values for $\rho = .6$ illustrating clearly that in spite of low selection ratios the criterion distribution is never far away from normality, i.e.,

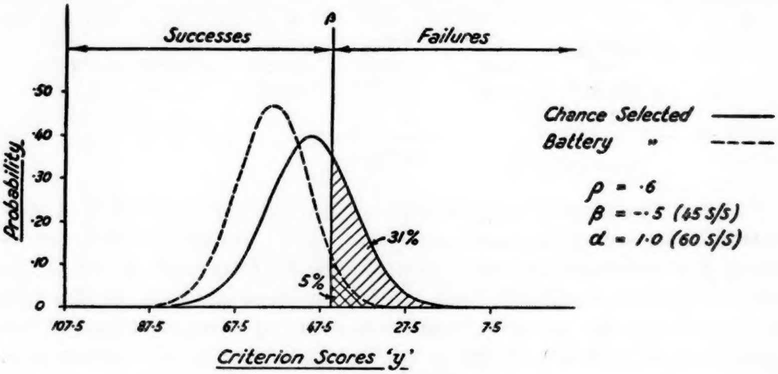

FIG. 5 QUALITY GAIN DIAGRAM

$\beta_1 \cong 0$ and $\beta_2 \cong 3$. It also shows that the selected group does not become very much more homogeneous than an unselected group would be because its standard deviation of true ability shrinks only a little even for strong curtailment on the predictor variable.

Such deviations from normality as exist could be detected neither by a χ^2 test nor by a significance test for the betas if our selected group is $N \cong 1000$. Only for correlation coefficients far in excess of the current ones would the criterion distribution of the selectees become markedly skew and peaked.

The Quality-Gain Diagram is obtained by plotting the probability distributions of true ability for the unselected and selected group. The shift of distributions against each other depicts graphically how much is gained by selecting a given number of applicants by aptitude testing in comparison to chance selection. The areas lopped off by the vertical axis through the cutting score β on the criterion variable give the wastage rates of the selectees for either method of selection. The difference of these areas is the efficiency index H. A Quality-Gain Diagram has been drawn for $\rho = .6$, $\alpha = 1.0$, $\beta = -0.5$ in Figure 5. The scale unit used is again the normalized standard score, S/S.

TABLE 1

Predictor Cut-Off α	Selection Ratio $\Phi(\alpha)$	Standard Deviation of Predictor Scores σ_x	Standard Deviation of Criterion Scores σ_y	Skewness of Criterion Scores $_y\beta_1$	Kurtosis of Criterion Scores $_y\beta_2$	Mean of Criterion Scores M_1'
$-\infty$	1	1	1	0	3	0
0	.5000	.6028	.8780	.00484	3.2503	.4787
+1	.1586	.4462	.8436	.00177	3.2026	.9151
+2	.0228	.3380	.8253	.00052	3.1095	1.4239
$+\infty$	0	0	.8	0	3	∞

$\rho = .6$

(*3*) *The Cost-Utility Diagram*: Although the proportion of successful selectees is the most important figure in any selection procedure, the employer undoubtedly will also be interested in the number of rejected applicants who had the required abilities for the job. In fact from the general *manpower point of view* selection of the capable and rejection of the incapable candidates is only achieved at the cost of accepting some applicants who do not make the grade and turning down quite a few others who would have proved successful if selected.

Berkson (7) has introduced the concepts of "cost" and "utility" of a test. To the author it seems almost impossible to give a general formula for balancing utility against cost, as circumstances differ so much from one selection program to another. For example, we may not mind rejecting many capable applicants for the vacancy of a train driver as long as the one we select is a good one. On the other hand, the number of rejected capable candidates for pilot training would be of major consequence in a national emergency. We may, however, show in an unambiguous manner how battery selection divides the applicant group into four distinct classes once α, β and ρ are determined.

By graphing the true ability distribution of the complete population of applicants and the distribution of the selected candidates, we arrive at the "Cost-Utility Diagram," this name being borrowed from Berkson. It is to be noted that for this particular purpose the criterion distribution of selectees is to be taken as

$$\psi(y)\,dy = \Phi(\alpha)\,\phi(y)\,dy = \mathrm{erf}\,(y)\,\pi(y)\,dy\,, \tag{20}$$

making the ratio of the areas under the two curves equal to the selection ratio. The distributions are shifted against each other similar to the displacement in the "Quality-Gain Diagram." The Vertical axis through the cutting score β on the true ability scale divides the area

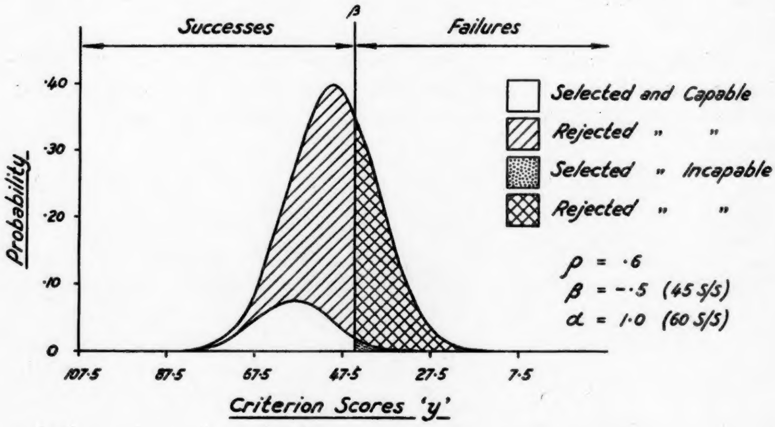

FIG. 6 COST-UTILITY DIAGRAM

under the total applicant distribution into four segments being proportional to

(1) the number of capable applicants selected by battery,
(2) the number of capable applicants rejected by battery,
(3) the number of incapable applicants selected by battery, and
(4) the number of incapable applicants rejected by battery.

In Figure 6 a Cost-Utility Diagram is shown for $\alpha = 1.0$, $\beta = -0.5$, $\rho = .6$.

It is hoped that the "Cost-Utility Diagram" will prove of value to the selector in putting the usefulness of a test battery across to management. Although it complies with rigorous scientific criteria, it does not require the understanding of a correlation coefficient or the explanation of probability. In spite of its simplicity it gives all the vital information in which the *employer* is interested.

II. *The Statistical Estimation*

(*a*) *The Confidence Belt of the Operating Characteristics*

The mathematical models of operating characteristics can be tested against actual data only if we are in a position to calculate the standard errors of their estimates. The two operating characteristics described in the previous section are analytically identical. We shall derive the equation for the probability limits of the S.O.C. curve. The whole argument is also applicable to the A.O.C. curve provided we interchange the variables and replace β by α.

Formulas derived previously were all given in standard measures. In practice we usually deal with raw test and criterion scores. Consequently we have to rewrite the basic equation (6) as

$$\pi(X) = \Phi\left\{\frac{1}{\sqrt{1-\rho^2}}\left[\left(\frac{L-\eta}{\sigma_y}\right) - \rho\left(\frac{x-\xi}{\sigma_x}\right)\right]\right\}, \qquad (21)$$

where

$$\frac{x-\xi}{\sigma_x} = X, \qquad (21a)$$

$$\frac{L-\eta}{\sigma_y} = \beta. \qquad (21b)$$

Equation (21a) amounts to a linear parametric transformation. Hence

$$\pi(X) = \pi(x).$$

We estimate $\pi(x)$ by

$$p(x) = \Phi\left\{\frac{1}{\sqrt{1-r^2}}\left[\left(\frac{L-\bar{y}}{s_y}\right) - r\left(\frac{x-\bar{x}}{s_x}\right)\right]\right\} \quad . \qquad (22)$$

We require the variance of $p(x)$. For large samples we have for the variance of a function involving several statistical estimates.

$$\mathrm{var}(\phi) = \Sigma\left\{\left(\frac{\partial\phi}{\partial z_j}\right)^2 \mathrm{var}(z_j)\right\} + \Sigma\left\{\frac{\partial\phi}{\partial z_j}\frac{\partial\phi}{\partial z_k}\mathrm{cov}(z_j, z_k)\right\} \quad , \qquad (23)$$

where z_j is the jth statistic involved. After finding the various partial derivatives in (22) and replacing the statistics by their expectations we have from (23)

$$\begin{aligned}\mathrm{var}\, p(x) = \frac{1}{(1-\rho^2)}&\left\{\mathrm{erf}\left(\frac{\beta-\rho X}{\sqrt{1-\rho^2}}\right)\right\}^2\left[\left(\frac{\rho\beta-X}{1-\rho^2}\right)^2 \mathrm{var}(r)\right.\\ &+\left(\frac{1}{\sigma_y}\right)^2 \mathrm{var}(\bar{y}) + \left(\frac{\rho}{\sigma_x}\right)^2 \mathrm{var}(\bar{x}) + \left(\frac{\beta}{\sigma_y}\right)^2 \mathrm{var}(s_y)\\ &+\left(\frac{\rho X}{\sigma_x}\right)^2 \mathrm{var}(s_x) - \frac{2\beta(\rho\beta-X)}{\sigma_y(1-\rho^2)}\mathrm{cov}(r,s_y) \qquad (24)\\ &+\frac{2\rho X(\rho\beta-X)}{\sigma_x(1-\rho^2)}\mathrm{cov}(r,s_x)\\ &\left.-\frac{2\rho}{\sigma_x\sigma_y}\mathrm{cov}(\bar{y},\bar{x}) - \frac{2\rho\beta X}{\sigma_x\sigma_y}\mathrm{cov}(s_y,s_x)\right] .\end{aligned}$$

In a normal bivariate distribution the estimates of the means are independent of the estimates of the standard deviations and the correlation coefficient. For this reason six covariances vanish in (23). Further we have in large sample theory for a normal bivariate distribution

$$\mathrm{var}(r) = \frac{(1-\rho^2)^2}{N},\ \mathrm{var}(\bar{y}) = \frac{\sigma_y^2}{N},\ \mathrm{var}(\bar{x}) = \frac{\sigma_x^2}{N},$$

$$\mathrm{var}(s_y) = \frac{\sigma_y^2}{2N},\ \mathrm{var}(s_x) = \frac{\sigma_x^2}{2N};$$

and

$$\operatorname{cov}(r,s_y)=\frac{\rho\,\sigma_y(1-\rho^2)}{2N},\ \operatorname{cov}(r,s_x)=\frac{\rho\,\sigma_x(1-\rho^2)}{2N},$$

$$\operatorname{cov}(\bar{y},\bar{x})=\frac{\rho\,\sigma_y\,\sigma_x}{N},\ \operatorname{cov}(s_y,s_x)=\frac{\rho^2\,\sigma_y\,\sigma_x}{2N}.$$

[See (8) page 38.] Substitution of these expressions into (24) gives

$$\begin{aligned}\operatorname{var} p(x)=\frac{1}{2N}\left(1+\frac{1}{1-\rho^2}\right)\left(X^2-\frac{2\rho\beta}{2-\rho^2}X+\frac{\beta^2+2(1-\rho^2)}{2-\rho^2}\right)&\\ \times\left\{\operatorname{erf}\left(\frac{\beta-\rho X}{\sqrt{1-\rho^2}}\right)\right\}^2.&\end{aligned}\tag{25}$$

Let us investigate equation (25) a little closer. For large positive or negative values of X the variance of the proportion of successes becomes very small; this means that we can make increasingly more accurate statements with respect to expectancies in the applicant population at the extremes of the predictor scale.

Differentiation of (25) and equating to zero gives

$$\begin{aligned}X^3-\frac{\beta(2+\rho^2)}{\rho(2-\rho^2)}X^2-\frac{2(1-\rho^2)-3\rho^2(\beta^2-\rho^2+1)}{\rho^2(2-\rho^2)}X&\\ -\frac{\beta(\beta^2-\rho^2+1)}{\rho(2-\rho^2)}=0.&\end{aligned}\tag{26}$$

Eq. (26) has one or three real roots according to

$$D=\left(\frac{q}{2}\right)^2-\left(\frac{p}{3}\right)^3\substack{\geq\\<}0,$$

where

$$\left.\begin{aligned}q&=-\frac{2(1-\rho^2)^2(2+3\rho^2)}{3\rho^3(2-\rho^2)^2}\beta-\frac{8(1-\rho^2)^2(2+7\rho^2)}{27\rho^3(2-\rho^2)^3}\beta^3,\\ p&=\frac{(1-\rho^2)(2-3\rho^2)}{\rho^2(2-\rho^2)}+\frac{2(1-\rho^2)(2-5\rho^2)}{3\rho^2(2-\rho^2)^2}\beta^2.\end{aligned}\right\}\tag{27}$$

If

$D>0$ we have one maximum of var $p(x)$,

$D = 0$ we have one maximum and one point of inflexion, with horizontal tangent,
$D < 0$ we have two maxima and one minimum.

The latter case is interesting as we have a constriction between the upper and lower probability limits somewhere between the extremes of the score range. In Figure 7 this type of probability limit is indicated for $\beta = 0$, $\rho = .6$, $N = 400$. From the population S.O.C. curve $\pi(x)$, $2.58\{\text{var } p(x)\}^{\frac{1}{2}}$ was laid off on either side. Where values above unity or below zero occurred they were taken as unity or zero. This must happen at the extremes of the scale as here, $p(x)$ cannot be distributed normally about $\pi(x)$ even for large N. Only once in a hundred times should an estimated S.O.C. curve based on one particular sample lie outside these limits.

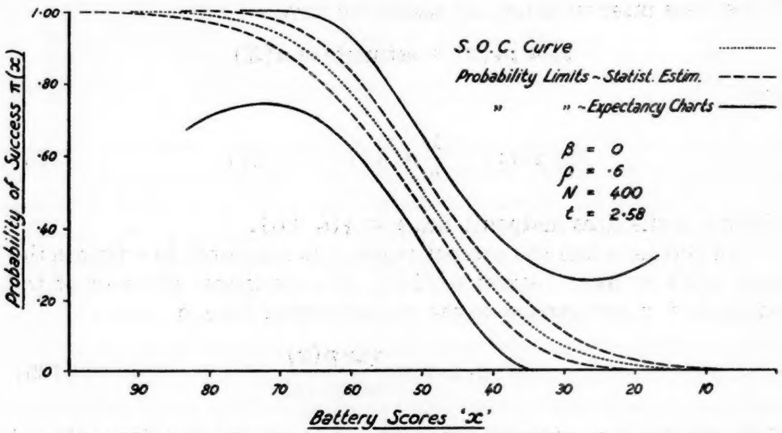

FIG. 7 PROBABILITY LIMITS OF TWO METHODS OF ESTIMATION

The above argument was based on the knowledge of the population S.O.C. curve. Usually we do not know it. Taking the S.O.C. curve based on the particular sample we deal with as the best estimate, we can lay off $2.58\{\text{var } p(x)\}^{\frac{1}{2}}$ from $p(x)$ and state that in repeated inferences of similar kind it will happen in the long run only once in a hundred times that a population S.O.C. curve $\pi(x)$ will lie outside limits derived in such a way. The area between the limits is the 99% confidence belt for estimating true expectancies $\pi(x)$.

The present usage of expectancy charts, seen from the viewpoint of estimation, is a very inefficient procedure. For each group of applicants scoring $l_1 \leq x \leq l_2$ on the predictor, an estimate of the population proportion of successes π, given in equation (7), is furnished by computing the proportion of observed successes p_E in each class interval. The variance of such an estimate is

$$\operatorname{var} p_E = \frac{1}{n}\pi(1-\pi)\,.$$

If the class interval is not too coarse we have

$$p_E \cong p_E(x) = \text{estimate of } \pi(X)$$

and

$$\operatorname{var} p_E(x) = \frac{1}{n}\pi(X)\,[1-\pi(X)]\,, \tag{28}$$

where x is the class midpoint, i.e., $x = \frac{1}{2}(l_1 + l_2)$.

In Bittner's and the present paper it is suggested to estimate the same $\pi(X)$ by $p(x)$ [equation (22)]. The statistical efficiency of the old method in comparison to the one advocated here is

$$\text{Eff.}\ p_E(x) = \frac{\operatorname{var} p(x)}{\operatorname{var} p_E(x)}\,. \tag{29}$$

If N, the total number of applicants, is large and the class interval width W (in standard measures) reasonably small, we may write without introducing any serious error in (28)

$$n \cong NW \operatorname{erf}(X)$$

and equation (29) becomes

$$\text{Eff.}\ p_E(x) = \frac{W}{2}\left(1+\frac{1}{1-\rho^2}\right)\left\{\operatorname{erf}\left(\frac{\beta}{\sqrt{1+\rho^2}}\right)\right\}^2 \frac{1}{\pi(X)\,[1-\pi(X)]}$$
$$\times\left(X^2-\frac{2\rho\beta}{2-\rho^2}X+\frac{\beta^2+2(1-\rho^2)}{2-\rho^2}\right)\operatorname{erf}\left(\frac{2\rho\beta-(1+\rho^2)X}{\sqrt{1-\rho^4}}\right). \tag{30}$$

From equation (30) the statistical efficiencies were calculated for some values of X for $\beta = 0$ and $\rho = .6$. From equations (28) and (25) the standard deviations of the estimates for $\pi(X)$ multiplied times $t = 2.58$ were also computed. N was taken as 400. All results are given in Table 2.

TABLE 2

Standard Measure X	Efficiency of $p_E(x)$ $(\beta=0, \rho=.6)$	$2.58\,\sigma[p_E(x)]$ $(\beta=0, \rho=.6,$ $N=400)$	$2.58\,\sigma[p(x)]$ $(\beta=0, \rho=.6,$ $N=400)$	$\pi(x)$ $(\beta=0, \rho=.6)$
0	.127	.145	.051	.500
.5	.140	.148	.054	.646
1.0	.143	.155	.059	.773
1.5	.100	.170	.054	.870
2.0	.044	.193	.041	.933
2.5	.013	.231	.027	.970
3.0	.002	.289	.012	.988

In the above table only positive deviations of X are given. As $\beta=0$, due to symmetry, identical numerical values are obtained for negative deviations. The statistical efficiencies for the estimator $p_E(x)$ are very low in comparison to estimator $p(x)$. For example, to arrive at the same accuracy of estimate for the population proportion of successes in the test score interval $-.25 \leq X \leq +.25$, the total number of applicants N has to be eight times greater using the current method of expectancy chart inference than when using equation (22).

The reason for the relative inefficiency of the current method of estimation is the small amount of information furnished by the number of applicants in one particular score interval only. On the other hand equation (22) makes effective use of all the information contained in the total sample N even when estimating the proportion of successes for one specific score interval.

The respective probability limits and the S.O.C. curve for $\beta=0$, $\rho=.6$ and $N=400$ are drawn in Figure 7. The graph illustrates strikingly how much may be gained by using a statistically efficient method of estimation.

(b) *The Standard Error of the Estimated Mean Criterion Score of Selectees*

The standard error of the estimated mean criterion score of selectees has been discussed in Jarrett's paper (6). The formula given is, as the author points out, a rough approximation as it does not take into account sampling fluctuations of the correlation coefficient and the standard deviations. Furthermore, it refers to an estimate based on $N\pi_s$, the number of selectees, only. Consequently a great deal of information is sacrificed as, by such a method of estimation, we do not make use of the data furnished by $N(1-\pi_s)$ applicants. In the following, the large sample variance of an estimate based on the total number of applicants is derived.

It has been mentioned previously that the cutting score a on the predictor variable is dependent on the selection ratio π_s. Theoretically we ought to find a from the parametric relation

$$\pi_s = \Phi(a),$$

and the raw predictor cutting score from

$$\lambda = a\,\sigma_x + \xi\,.$$

The practical approach, however, is slightly different. From the mean and standard deviation of the first sample we estimate

$$\lambda' = a\,s_{x_1} + \bar{x}_1\,.$$

Having thus determined λ' we keep it fixed in future applications of the test or battery. λ', which originally was an estimate of λ, becomes a parameter of its own, but

$$a' = \frac{\lambda' - \xi}{\sigma_x} \neq \frac{\lambda - \xi}{\sigma_x} = \alpha.$$

Hence in practice the selection ratio obtained will be different from π_s when selection takes place at a fixed cut-off λ'. The difference will not be very pronounced if the total number of applicants N on which $\bar{x}_1$ and s_{x_1} are based is large. The raw score criterion mean of the selectees in the population of applicants is

$$\gamma = \frac{\operatorname{erf}\left(\dfrac{\lambda' - \xi}{\sigma_x}\right)}{\Phi\left(\dfrac{\lambda' - \xi}{\sigma_x}\right)}\,\rho\,\sigma_y + \eta\,, \tag{31}$$

and is estimated by

$$g = \frac{\operatorname{erf}\left(\dfrac{\lambda' - \bar{x}}{s_x}\right)}{\Phi\left(\dfrac{\lambda' - \bar{x}}{s_x}\right)}\,r s_y + \bar{y}\,. \tag{32}$$

The variance of g may be derived in the same way as var $p(x)$ using equation (23) and the same variances and covariances for the individual five statistics involved.

We finally have for the variance of the estimate of the mean criterion score of selectees

$$\mathbf{var}(g) = \left\{ 1 + \left(1 - \frac{\rho^2}{2}\right) m_1'^2 + \rho^2 \left(1 + \frac{a'^2}{2}\right)\left(m_1' - a'\right)^2 m_1'^2 \right.$$

$$\left. - \rho^2 \left(2 + a' m_1'\right)\left(m_1' - a'\right) m_1' \right\} \frac{\sigma_y^2}{N}, \qquad (33)$$

where m_1' and a' are the mean and cut-off point of the truncated normal distribution expressed in standard measures. For the special case of $a' = 0$ and $\rho = .6$, the variance of g is only 6/10 of the variance given by Jarrett in spite of having taken account of the sampling fluctuations of the five statistics involved. The reason for this greater efficiency is, as already mentioned, the larger number of applicants contributing to the estimation process.

(*c*) *The Standard Error of the Predicted Number of Successful Selectees at Score Level* x

After testing it is often imperative to make some kind of a forecast with respect to the selected applicants which may be verified from subsequent validations. In the case of an individual we cannot say more than that his chance of success is $\pi(x)$. We cannot check his *a priori* chance from a validation study. However, we may check the *a priori* chance of the total group of selectees by comparing the actual proportion of successes with π_2 of equation (8a).

The total group of selectees is in itself rather variable with respect to test scores and in many applied situations the raw predictor cutting score λ' shifts in time due to the needs of the employer and the abundance or shortage of labour. It is, therefore, more advantageous to compare the probabilities of groups of selectees falling within a predictor score range $l_1 \leq x \leq l_2$ with proportions of successes at the follow-up stage. In fact, such a procedure gives real meaning to the probability of success of an individual as it implies that of E_x persons, all attaining the same battery score x as he does, $E_x \pi(x)$, in the long run, will prove successful on the job. The standard error of the predicted number of successful selectees is

$$\sigma\{E_x p_E(x)\} = \sqrt{E_x \pi(x)\,[1 - \pi(x)]}\,. \qquad (34)$$

For a group of E_x chosen applicants all having test scores $l_1 \leq x \leq l_2$ we expect with confidence of $P = .99$,

$$E_x \pi(x) \pm 2.58 \sqrt{E_x \pi(x)\,[1 - \pi(x)]} \qquad (35)$$

to prove successful in the occupation for which they were selected. Eq. (35) is an approximation since for small E_x the binomial distribution deviates from the normal. $\pi(x)$ has to be estimated from $p(x)$ of equation (22). For small and moderate E_x the standard error of $p_E(x)$ is much larger than that of $p(x)$. Even for comparatively large numbers of applicants N, the number per array E_x is moderate or smallish. Hence the substitution of

$$p(x) \cong \pi(x)$$

in (35) does not cause serious errors.

Throughout the paper we have tacitly assumed that we deal with a statistically uniform universe. If any one of the means, standard deviations, or covariances change in time, selection becomes useless, as all concepts of probability of success, efficiency index, etc. are affected. It is advisable to install a statistical quality control program for predictor and criterion variables in order to get advance information on time shifts of population parameters.

III. Application

The foregoing principles were applied to data collected during World War II. A battery of eight tests was given to a group of artisan trainees of the South African Air Force before entering the training school. The whole group was permitted to go through the course at the end of which an objective trade test was given incorporating the subject matter covered during the training period. Altogether there were 287 trainees for whom both battery and criterion scores were available.

In order to apply the formulas derived in this investigation it is necessary that the data conform with the assumption of a bivariate normal population. Both raw battery and raw criterion scores of the above group were non-normal. The first step then was to normalize the marginal distributions with the help of a graphical procedure* which simultaneously standardized the distributions by making the means and standard deviations of the sample approximately equal to 50 and 10, respectively. The normalized standard score distributions of the group are given in Table 3 together with the χ^2 tests for normality. Although the graphical method used for normalizing the distributions constitutes a legitimate estimation procedure, it is not known how many degrees of freedom should be deducted from the

*A description of this method is considered to be beyond the scope of this study.

TABLE 3*

Standard Scores	Expected	Battery Scores Observed	Criterion Scores Observed
80 – 85	1.8	1	–
75 – 80		1	1
70 – 75	4.7	5	5
65 – 70	12.6	10	10
60 – 65	26.4	26	31
55 – 60	43.0	39	45
50 – 55	55.0	55	56
45 – 50	55.0	53	48
40 – 45	43.0	44	47
35 – 40	26.4	31	27
30 – 35	12.6	13	11
25 – 30	4.7	8	5
20 – 25	1.8	–	1
15 – 20		1	–
N	287.0	287	287
$\chi^2_{[8]}$	—	1.947	2.924
$P_{[5]}$	—	.85	.71
Mean	50.0	49.3	50.1
S.D.	10.0	10.2	9.8

*The brackets in the table indicate the groupings for the χ^2 test.

total cell numbers. For matters of expedience it is suggested to use the customary three constraints. If we do this, we must not interpret the χ^2 test with the help of a probability statement as we ordinarily do, but we should look at it as a measure of success for our normalization procedure.

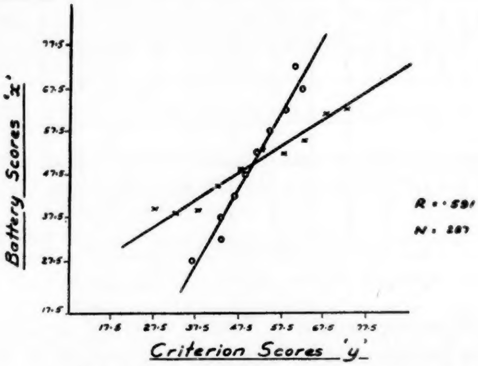

FIG. 8. OBSERVED REGRESSION LINES

Although the marginal scores appear to have been normalized satisfactorily according to Table 3, it does not follow that the bivariate distribution is normal. The next step consists of setting up a bivarate table where each subject's battery and criterion raw scores are converted into a pair of normalized standard scores (Table 4). Linearity of regression may then be tested by plotting the observed regression lines (Figure 8).

TABLE 4

Criterion Scores y

Battery Scores x	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65	65-70	70-75	75-80	80-85	Totals E_x	Observed $p_c(x)$
80-85											1				1	
75-80										1					1	1.000
70-75								1	3				1		5	
65-70								1	2	2	3	2			10	1.000
60-65						1	1	8	5	8	1	2			26	.962
55-60						3	6	15	7	5	3				39	.923
50-55			1	1	1	8	14	12	10	6	1	1			55	.800
45-50			1		6	10	14	10	8	4					53	.679
40-45				4	6	11	8	7	4	3	1				44	.523
35-40			1	3	8	10	3	1	3	2					31	.290
30-35			1	1	3	3	2	1	2						13	.385
25-30		1	1	2	2	1			1						8	
20-25															-	.111
15-20					1										1	
Totals E_y	-	1	5	11	27	47	48	56	45	31	10	5	1	-	287	.683
Observed $p_c(y)$		.000		.000	000	.021	.021	.179	.222	.355	.500		.833		.150	

λ'

L

The observed array means in Figure 8 do not deviate more from linear regression than would be expected in a sample of $N = 287$. We may now assume that the data of Table 4 conform with the hypothesis of a normal bivariate population. In standard scores we have

$$\begin{array}{ll} \bar{y} = 50.1 \quad , & \bar{x} = 49.3 \, , \\ s_y = \ 9.8 \quad , & s_x = 10.2 \, , \\ R = \ \ .591 \, , & L = 45.0 \, . \end{array}$$

Substitution of these values into equations (22) will yield the required best estimate of the S.O.C. curve. Prior to the derivation of the formulas in Part II, a detailed numerical analysis had been undertaken for

$$\begin{array}{ll} \bar{y} = 50.00 \quad , & \bar{x} = 50.00 \, , \\ s_y = 10.00 \quad , & s_x = 10.00 \, , \\ R = \ \ .5906 \, , & L = 45.00 \, . \end{array}$$

As these values are not radically different from the above correct sample values, the author may be forgiven for using the latter set for purpose of illustration rather than the former. The equation of the selector's operating characteristic becomes

$$p(x) = \Phi(3.03979 - .073188x),$$

where x is measured in standard scores. $p(x)$ is evaluated in Table 5 and graphed together with the observations (based on Table 4) in Figure 9. The 99% confidence belt of the S.O.C. curve as derived from an expectancy chart may be obtained from Hald's Tables **(9)**. For the estimation by equation (22) we have to lay off 2.58[var $p(x)]^{\frac{1}{2}}$ from $p(x)$ for a corresponding 99% confidence belt. The expression for var $p(x)$ (equation 25) contains parameters ρ and β which may be estimated by R and $\frac{L - \bar{y}}{s_y}$. Substitution of the sample values gives

$$\operatorname{var} p(x) = .00442(X^2 + .358X + .941)\{\operatorname{erf}(-.619 - .731X)\}^2.$$

X is estimated from $\frac{x - \bar{x}}{s_x}$ where the x's are the usual standard scores. The 99% upper and lower confidence limits for the expectancy chart method of estimation are given in the columns $\bar{\theta}$ and $\underline{\theta}$ in Table 5 and the limits for the statistical method in columns $\bar{\Delta}$ and $\underline{\Delta}$. A glance at Fig. 9 shows how much narrower the confidence belt of the statistical estimation is. Incidentally, this particular battery furnishes an example for a confidence belt without a constriction in the centre, i.e., $D > 0$.

TABLE 5

Standard Scores 'x'	$3{\cdot}03979 - {\cdot}073188x$	Estimated $p(x)$	Observed $p_E(x)$	E_x	$E_x p_E(x)$	$\overline{\theta}$	$\underline{\theta}$	$2{\cdot}58s[p(x)]$	$\overline{\Delta}$	$\underline{\Delta}$
2.5	2.856	.002	—	—	—	—	—	.005	.007	.000
7.5	2.491	.006						.013	.019	.000
12.5	2.125	.017						.027	.044	.000
17.5	1.759	.039	.111	9	1	.585	.001	.047	.086	.000
22.5	1.393	.082						.071	.153	.011
27.5	1.027	.152						.092	.244	.060
32.5	.661	.254	.385	13	5	.755	.094	.100	.354	.154
37.5	.295	.384	.290	31	9	.536	.110	.094	.478	.290
42.5	– .071	.528	.523	44	23	.716	.324	.076	.604	.452
47.5	– .437	.669	.679	53	36	.830	.494	.060	.729	.609
52.5	– .802	.789	.800	55	44	.917	.628	.052	.841	.737
57.5	–1.168	.879	.923	39	36	.991	.745	.046	.925	.833
62.5	–1.534	.938	.962	26	25	1.000	.747	.036	.974	.902
67.5	–1.900	.971	1.000	10	10	1.000	.589	.024	.995	.947
72.5	–2.266	.988	1.000	7	7	1.000	.469	.014	1.000	.974
77.5	–2.632	.996						.007	1.000	.989
82.5	–2.998	.999						.003	1.000	.996

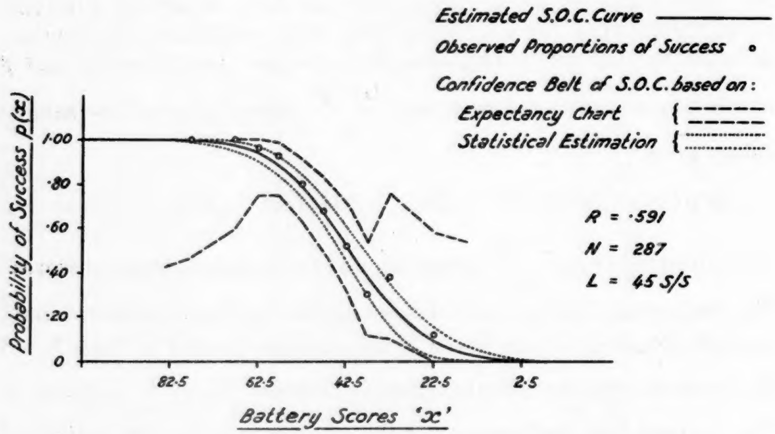

FIG. 9 SELECTOR'S OPERATING CHARACTERISTIC

TABLE 6

Standard Scores y	Observed $p_E(y)$	Estimated $p(y)$	Estimated $\text{erf}\left(\frac{y-\bar{y}}{s_y}\right)$	Estimated Ordinates $\left[\Phi\left(\frac{\lambda'-\bar{x}}{s_x}\right)\right]^{-1} p(y)\,\text{erf}\left(\frac{y-\bar{y}}{s_y}\right)$	Class Interval Probabilities		$p(y)\,\text{erf}\left(\frac{y-\bar{y}}{s_y}\right)$
17.5	.000	.000	.0020	.000	.000		.000
22.5		.001	.0091	.000	.000		.000
27.5		.002	.0317	.000	.000		.000
32.5	.000	.006	.0863	.003	.002		.001
37.5	.000	.016	.1826	.018	.010		.003
42.5	.021	.037	.3011	.070	.036		.011
47.5	.021	.078	.3867	.190	.096	$p_2 =$ .925	.030
52.5	.179	.146	.3867	.356	.177		.056
57.5	.222	.245	.3011	.465	.229		.074
62.5	.355	.373	.1826	.429	.212		.068
67.5	.500	.517	.0863	.281	.140		.045
72.5	.833	.658	.0317	.131	.067		.021
77.5		.780	.0091	.045	.024		.007
82.5		.873	.0020	.011	.006		.002
87.5	—	.934	.0004	.002	.001		.000
92.5		.969	.0001	.000	.000		.000
97.5		.987	.0000	.000	.000		.000
102.5		.995	.0000	.000	.000		.000
107.5		.999	.0000	.000	.000		.000

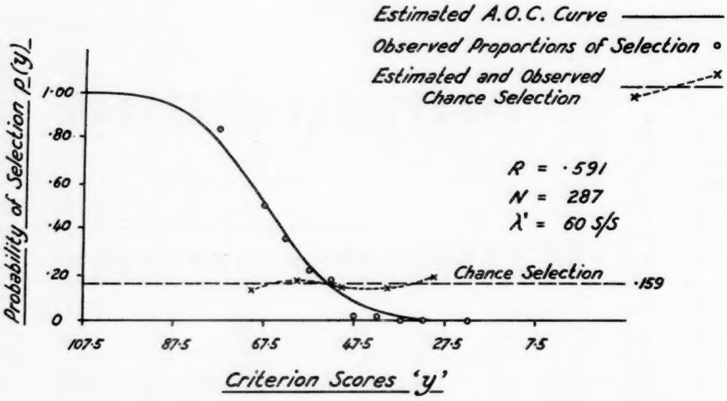

FIG. 10 APPLICANT'S OPERATING CHARACTERISTIC

The Applicant's Operating Characteristic was calculated from equation (22) for a battery cutting score of $\lambda' = 60{\cdot}00\ S/S$. It is

$$p(y) = \Phi(4.89862 - .073188y).$$

Estimated and observed values are shown in Table 6 and Figure 10. According to Table 4 there were 43 trainees who exceeded the battery cutting score $\lambda' = 60.0$. These 43 would have been chosen in case of an actual battery selection. In order to show what happens if 43 applicants were selected by a chance procedure, numbers from 1 to 287

TABLE 7

Criterion Score Range $y(S/S)$	Observed Frequency $N = 43$	Observed Frequency E_y $N = 287$	Observed Proportion of Selection	Expected Proportion of Selection	Observed Proportionate Frequencies $N = 43$ Chance Selection	Battery Selection
75–	–	1	—	.16	.000	.023
65–75	2	15	.13	.16	.047	.209
55–65	13	76	.17	.16	.302	.489
45–55	15	104	.14	.16	.349	.256
35–45	10	74	.14	.16	.232	.023
25–35	3	16	.19	.16	.070	.000
–25	–	1	—	.16	.000	.000
Totals	43	287	.150	.159	1.000	1.000

were assigned to each of the trainees. With the help of a random sampling number table, 43 numbers were drawn at random and the battery and criterion scores of the individuals to whom the drawn numbers pertained were noted. The result of this experiment is given in Table 7 and plotted in Figure 10. It shows again experimentally that the probability of selection is the same for all individuals irrespective of their true abilities if the selection procedure follows a chance pattern.

We now proceed to set up the Quality-Gain Diagram. The true ability (criterion) distribution of the unselected population of applicants (and also of the chance selected trainees) is

$$\operatorname{erf}(Y)dY = \frac{1}{\sqrt{2\pi}} e^{-\frac{Y^2}{2}} dY,$$

with Y in standard measures. As

$$Y = \frac{y - \eta}{\sigma_y} ,$$

we estimate Y by $\frac{y - \bar{y}}{s_y}$ so that we may find the ordinates of the criterion distribution of the chance selected group from

$$\operatorname{erf}\left(\frac{y - \bar{y}}{s_y}\right)$$

where the y's are observed standard scores. These ordinates are worked out in the fourth column of Table 6 and graphed as the solid curve in Figure 11. The criterion distribution of the selectees was given in equation (11) as

$$\phi(Y)dY = \frac{1}{\Phi(a)} \operatorname{erf}(Y)\pi(Y)dY.$$

Its ordinates may be estimated from

$$\left[\Phi\left(\frac{\lambda' - \bar{x}}{s_x}\right)\right]^{-1} \operatorname{erf}\left(\frac{y - \bar{y}}{s_y}\right) p(y).$$

Numerical values are given in the fifth column of Table 6 and are shown as the broken curve in Figure 11. In order to compare the theoretical distributions with the observations the observed propor-

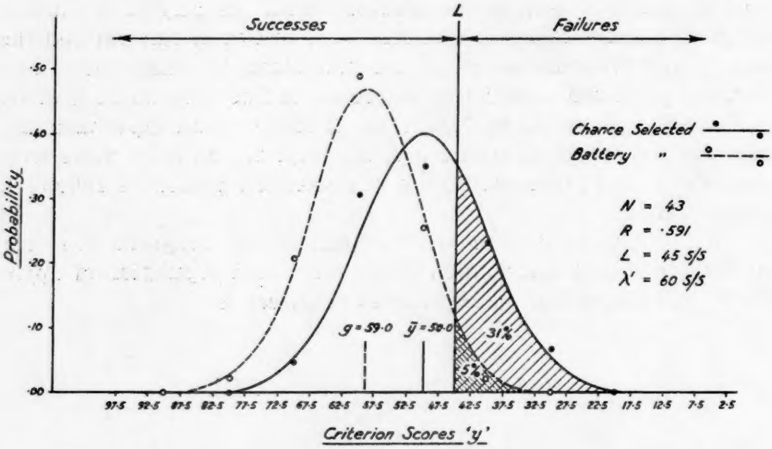

FIG. 11 QUALITY GAIN DIAGRAM

tionate frequencies are listed in the last two columns of Table 7. The sample size for both methods is $N = 43$.

The chance selection figures are based on the observations of the drawing experiment (Table 7) and the battery selection data originate from the criterion distribution of the selectees in Table 4. All observations are plotted in Figure 11 and show satisfactory agreement with the estimated distributions. Previously it had been mentioned that formulas (7) and (8a) could be evaluated numerically by a mechanical quadrature. Formula (8a) may be written

$$\pi_2 = [\Phi(\alpha)]^{-1} \int_{\beta}^{+\infty} \operatorname{erf}(Y)\pi(Y)dY$$

and its statistical estimate

$$p_2 = \left[\Phi \left(\frac{\lambda' - \bar{x}}{s_x} \right) \right]^{-1} \int_{\frac{L-\bar{y}}{s_y}}^{+\infty} p(y) \operatorname{erf}\left(\frac{y-\bar{y}}{s_y} \right) d\left(\frac{y-\bar{y}}{s_y} \right) .$$

A comparison of the heading of the fifth column in Table 6 and the function to be integrated in this latter expression shows the two to be identical. We can, therefore, effect the integration by first switching from ordinates to areas with the help of a quadrature formula and then adding the partial areas from the upper tail to the limit $\frac{L-\bar{y}}{s_y} = 45.0\ S/S$. The ordinates were calculated for the midpoints of the class intervals and we can use the simple quadrature formula

$$\int_{-\frac{1}{2}}^{+\frac{1}{2}} f(x)\,dx = \frac{W}{24}\,(y_{-1} + 22\,y_0 + y_{+1})\,,$$

where W is the width of the class interval in standard measure. The resulting areas (which represent the class interval probabilities of the estimated criterion distribution of selectees) are shown in the sixth column of Table 6. Addition of the probabilities above 45 S/S finally gives the estimated proportion of successful selectees

$$p_2 = .952\,.$$

Computation of p_2 with the help of Pearson's Tables is complicated but more exact. It leads to

$$p_2 = .9517\,.$$

The estimated course wastage rate is

$$W.R. = 100\,(1-p_2) = 4.8\%\,.$$

For a chance selection we have from equation (8b)

$$\pi_1 = \Phi(\beta)$$

and its estimate

$$p_1 = \Phi\left(\frac{L-\bar{y}}{s_y}\right) = \cdot 691.$$

The corresponding wastage rate is

$$W.R. = 100\,(1-p_1) = 30.9\%\,.$$

The course wastage rates are shown as the shaded portions in Figure 11. For this particular example test selection reduces the course wastage rate based on chance procedure by more than five sixths.

The estimated efficiency index of the battery is from equation (8)

$$\hat{H} = 100\,(p_2 - p_1) = 26.1\%\,.$$

Consequently, by means of this test battery, we expect to gain over chance procedure 26 successful trainees in every 100 selected.

The estimated mean criterion score of the chance selected trainees is $\bar{y} = 50.0\ S/S$ and that of the battery selected group is given by equation (32)

$$g = \frac{\operatorname{erf}\left(\dfrac{\lambda' - \bar{x}}{s_x}\right)}{\Phi\left(\dfrac{\lambda' - \bar{x}}{s_x}\right)}\, r\, s_y + \bar{y} = 59{\cdot}0\ S/S\,.$$

Hence the average quality of the battery selected group represents an improvement of 9 standard scores over a chance selected group.

Table 8 affords a comparison between some expected and observed quantities used to describe various aspects of selection. The expected values are based on statistical estimates derived from the total sample of $N = 287$, whereas the observed values are based on samples of $N = 43$, the number of trainees actually selected, and shown in Table 7.

TABLE 8

	Battery Selection		Chance Selection	
Parameter	Statistical Estimate $N = 287$	Observed $N = 43$	Statistical Estimate $N = 287$	Observed $N = 43$
Percentage of successful selectees $100\,\pi_2$ & $100\,\pi_1$	95%	98%	69%	70%
Wastage Rates	5%	2%	31%	30%
Efficiency Index H	26%	28%	—	—
Criterion Mean Scores γ & η	59.0	59.5	50.0	50.2

The standard error of $\bar{y}$, based on the total sample of $N = 287$, is

$$\sigma(\bar{y}) = \frac{s_y}{\sqrt{N}} = 0 \cdot 6\ S/S\,,$$

and that of g, also based on $N = 287$ and computed from equation (35),

$$\sigma(g) = 0 \cdot 9\ S/S\,.$$

ting score L without, however, altering the selection ratio and for

Suppose now that we were permitted to shift the criterion cut-that matter λ'. For maximum efficiency index we have from equation (10)

$$\beta_{\max} = \left(\frac{1 - \sqrt{1 - \rho^2}}{\rho} \right) \alpha .$$

The estimate for $\beta_{\max}$ is

$$b_{\max} = \left(\frac{1 - \sqrt{1 - R^2}}{R} \right) \left(\frac{\lambda' - \bar{x}}{s_x} \right) = .327$$

and

$$L_{\max} = s_y \, b_{\max} + \bar{y} = 53{\cdot}3 \; S/S \, .$$

For such a cutting score the Efficiency Index H is approximately 38% which is considerably higher than the one we had found for $L = 45$ S/S. Just to show that this upward difference in efficiency is also present in the sample of 287 trainees let us put L at 55 S/S which is not far removed from $L_{\max}$. From Table 4 we see that 31 out of the 43 selected would have had criterion scores $\geqslant$ 55 S/S and

$$100 p_2' = \frac{31}{43} = 72\% .$$

If no selection had taken place, 92 of the 287 trainees would have had criterion scores $\geqslant$ 55 S/S according to the same table and

$$100 p_1' = \frac{92}{287} = 32\% .$$

Consequently

$$\hat{H} = 100 \, (p_2' - p_1') = 40\% ,$$

which is 12% higher than the observed Efficiency Index for $L = 45$ S/S. (That this $\hat{H}$ exceeds the maximum of 38% is due to chance, as p_2' is based on $N = 43$ only).

In constructing the Cost-Utility Diagram we again use the ordinates of the estimated criterion distribution of the total applicant population shown in the fourth column of Table 6. The corresponding observations are given by the marginal criterion distribution of the total number of trainees in Table 4. In order to reduce the actual frequencies E_y to comparable ordinates, we have to divide the observed cell frequencies by a factor $N_1 W$ where $N_1 = 287$, the total sample number and $W = 0.5$, the width of the class interval in standard measures. The estimated and observed criterion distribution of all the trainees is plotted in Figure 12.

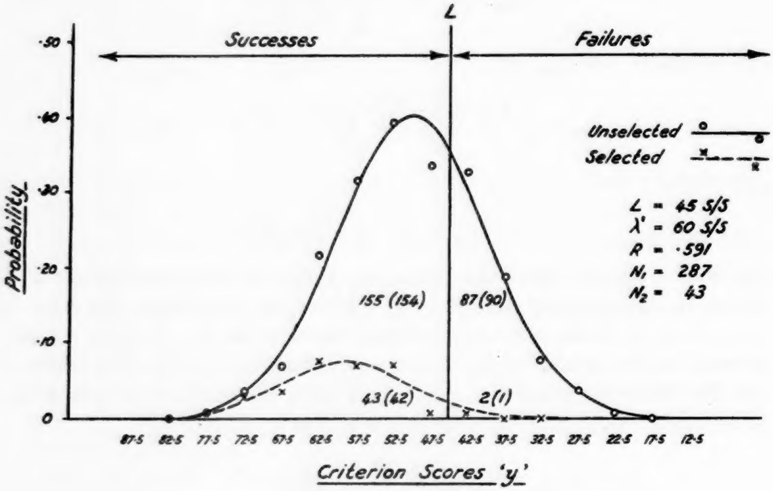

FIG. 12 COST-UTILITY DIAGRAM

For the purpose of the Cost-Utility Diagram the criterion distribution of the selectees is given by equation (20)

$$\Psi(Y)dY = \text{erf}(Y)\pi(Y)dY$$

and its ordinates are estimated from

$$\text{erf}\left(\frac{y-\bar{y}}{s_y}\right)p(y).$$

The numerical values are computed by multiplying figures in the fifth column of Table 6 with a factor $\Phi\left(\frac{\lambda'-\bar{x}}{s_x}\right)$. They are shown in the last column of the same table. To arrive at corresponding observed ordinates we first form the marginal criterion distribution of the $N_2 = 43$ selectees by adding all columns above the cutting score $\lambda' = 60$ S/S in Table 4. Next we divide individual cell frequencies by the factor N_1W which we used already before. Finally observed and estimated ordinates are plotted in the Cost-Utility Diagram (Fig. 12). The axis through $L = 45$ S/S divides the Cost Utility Diagram into four areas representing four groups of trainees. The expected numbers in these partitions may be calculated from the following fourfold Table 9(a).

TABLE 9

Number of capable trainees rejected $N(\pi_1 - \pi_0 \pi_2)$	Number of incapable trainees rejected $N(1 - \pi_1 - \pi_0 + \pi_0 \pi_2)$	Total Number of rejected trainees $N(1 - \pi_0)$	155 (154)	87 (90)	242 (244)
Number of capable trainees selected $N \pi_0 \pi_2$	Number of incapable trainees selected $N(\pi_0 - \pi_0 \pi_2)$	Total Number of selected trainees $N \pi_0$	43 (42)	2 (1)	45 (43)
Total number of capable trainees $N \pi_1$	Total number of incapable trainees $N(1 - \pi_1)$	Total number of trainees N		198 (196)	89 (91) 287 (287)
	(a)			(b)	

where

$$\pi_0 = \Phi\left(\frac{\lambda' - \xi}{\sigma_x}\right) = \Phi(a')$$

and π_1 and π_2 are as previously defined. The estimates for π_0, π_1 and π_2 were also computed before. They are

$$p_0 = .159\,,$$
$$p_1 = .691\,,$$
$$p_2 = .952\,.$$

For $N = 287$ we expect the four groups to be of magnitudes as indicated in Table 9(b). The actually observed frequencies derived from Table 4 are shown in parentheses. To the casual observer it may appear that the agreement between expected and observed frequencies in Table 9(b) and in general between all the expected and observed values mentioned in this investigation seems to be too good to be true. The reason for this is, of course, that we actually have six constraints operating by making the means, variances, covariance, and number of the sample equal to corresponding quantities in the expected bivariate distribution. Furthermore, it is not clear how many more degrees of freedom were lost in the initial normalization procedure applied to the raw score sample.

In Table 10 a final comparison between the expected and observed number of successful selectees is given. In addition upper and lower probability limits are shown. The numerical values are based on formula (35).

TABLE 10

Standard Scores x	E_x	$p(x)$	Expected Successes $E_x p(x)$	Observed Successes $E_x p_E(x)$	99% Probability Limits	
					Lower	Upper
80–85	1	.999	1.0	1	1*	1*
75–80	1	.996	1.0	1	1*	1*
70–75	5	.988	4.9	5	4*	5*
65–70	10	.971	9.7	10	8	10
60–65	26	.938	24.4	25	21	26
55–60	39	.879	34.3	36	29	39
50–55	55	.789	43.4	44	36	51
45–50	53	.669	35.5	36	27	44
40–45	44	.528	23.2	23	15	32
35–40	31	.384	11.9	9	5	19
30–35	13	.254	3.3	5	0	7
25–30	8	.152	1.2	1	0*	4*
20–25	–	.082	–	–	–	–
15–20	1	.039	0.0	0	0*	1*
Totals	287		193.8	196		

Probability limits marked with asterisks were computed from the accurate binomial distribution. Expected and observed frequencies of successful selectees are closer than the probability limits would suggest. This again is due to the fact that the theoretical probabilities of success $\pi(x)$ were estimated from this particular sample. In addition we now introduce new constraints by using the individual sample frequencies E_x in computing the expected number of successes. This also accounts for the discrepancy between the total number of expected successes in Table 10 (193.8) and 9(b) (198). A far more stringent test would be to make a comparison between expected and observed successes of an independent sample but using the estimates of the original.

However, the approach to subsequent follow-up studies as indicated in Table 10 has many advantages. After the initial validation we generally deal with truncated samples owing to the fact that selection has already taken place. Cutting scores very frequently are adjusted to the needs of the day. Often a short-time intake is not representative of the long-time distribution of scores above the cut-off point. Yet we need not be disturbed about all these factors if we make predictions in terms of probable successes and validate by having these predictions come true within the previously stated probability limits.

REFERENCES

1. McClelland, W. Selection for secondary education. London: Univ. of London Press, 1942.
2. Bittner, R. H., and Wilder, C. E. Expectancy tables: a method of interpreting correlation coefficients. *J. exp. Educ.*, 1946, 14, 245-252.
3. Bingham, W. V. Great expectations. *Personnel Psych.*, 1949, 2, 397-404.
4. Taylor, H. C., and Russell, J. T. The relationship of validity coefficients in the practical effectiveness of tests in selection: Tables and Discussions. *J. appl. Psych.*, 1939, 23, 565-578.
5. Brogden, H. E. When testing pays off. *Personnel Psych.*, 1949, 2, 170-183.
6. Jarrett, R. F. Per cent increase in output of selected personnel as an index of test efficiency. *J. appl. Psych.*, 1948, 32, 135-145.
7. Berkson, J. "Cost-utility" as a measure of the efficiency of a test. *J. Amer. statist. Ass.*, 1947, 42, 246-55.
8. Kendall, M. G. The advanced theory of statistics, Vol. II. London, Charles Griffin and Co., 1946.
9. Hald, A. Statistiske metoder, Tabel-Og formelsamling. Copenhagen, 1948, pp. 54-55.

Manuscript received 12/20/50.

PSYCHOMETRIKA—VOL. 17, NO. 1
MARCH, 1952

NOTE ON THE COMPUTATION OF BISERIAL CORRELATIONS IN ITEM ANALYSIS

LAURENCE SIEGEL
STATE COLLEGE OF WASHINGTON
AND
EDWARD E. CURETON
UNIVERSITY OF TENNESSEE

A rapid method is described for machine computation of biserial correlations in item analysis with several criteria. This method has been found to yield biserial correlations from punched IBM cards at the rate of about 41 per hour.

Biserial correlation techniques are used extensively in determining the discriminating powers of test items. Dubois, Dunlap, and Royer (1, 2, 3) have described methods for the rapid calculation of biserial correlations by means of nomographs or IBM equipment. The most recent of these methods (1) requires that each item be coded in a separate column of a standard 80-column IBM card.

The present method was devised in connection with a problem which required the correlating of 483 item-responses with each of eleven criteria ($N = 597$). Each item was actually an alternative of one of 97 multiple-choice questions in a biographical inventory. Each item-response is of course dichotomous; the individual either does or does not mark it. Ten item-responses were punched in each column of the IBM card, punching only if the item was marked.* The criterion scores were assumed to represent continuous variables, but each of them was coded in ten categories (0-9) so it could be punched as one digit in a single column of the card. The card design was as follows:

Columns 1-4	Subject identification number
Columns 5-53	Item responses
Columns 60-70	Criterion scores

The formula for the biserial correlation may be written:

*An alternative procedure would have been to code each item in a separate column and to use the reproducer to punch the criterion scores into multiple cards.

$$r_{bi} = \frac{m - M}{\sigma} \cdot \frac{p}{z}. \tag{1}$$

where

m is the mean criterion score of the group selecting a given item;

M is the mean criterion score of the total group;

σ is the standard deviation of the criterion scores of the total group;

p is the proportion of the total group who marked the given item; and

z is the ordinate of the unit normal distribution corresponding to the tail-area p.

Let

N be the number in the total group,

n the number marking the given item,

$\sum$ a summation from 1 to N, and

S a summation from 1 to n.

Then

$$p = \frac{n}{N} \tag{2}$$

$$m = \frac{SX}{n} \tag{3}$$

$$M = \frac{\sum X}{N} \tag{4}$$

$$\sigma^2 = \frac{\sum X^2 - M \sum X}{N - 1}. \tag{5}$$

Formula (1) may be rewritten in the form,

$$r_{bi} = (1/N\sigma)\,(1/z)\,(SX) - (p/z)\,(M/\sigma). \tag{6}$$

A single sort on each criterion provides a frequency distribution from which M and σ may be computed by (4) and (5). The values of $(1/N\sigma)$ and (M/σ) are then constant for all item-correlations with a given criterion.

The cards are then sorted on each item position. (Wherever a question has mutually exclusive alternatives, a single sort provides subgroups of cards for each of the alternatives.) The subgroup of cards for those who marked each given item is then run through the tabulator, which is used to accumulate the eleven sets of criterion scores. The card count is the value of n for the item and the eleven totals are the values of SX for the eleven criteria.

Since N remains constant throughout, a special table giving values of $(1/z)$ and (p/z) directly for all possible values of n may be prepared from standard tables of the normal distribution. Dunlap (2) has already tabulated the values of (p/z) with argument p. The values of $(1/z)$ and (p/z) are constant for each item in all criteria; only SX is different for every item and criterion.

The sorting and tabulating for 5,313 correlations required 85 hours. Less sorting time would have been required if all questions had had mutually exclusive alternatives. Subsequent computations on a desk calculator produced biserial correlations at an average rate of about 120 per hour. Thus the data were converted from punched IBM cards to biserial correlations at the rate of about 41 per hour.

REFERENCES

1. Dubois, Philip H. Note on the computation of biserial r in item validation. *Psychometrika*, 1942, 7, 143-146.
2. Dunlap, Jack W. Note on computation of biserial correlations in item evaluation. *Psychometrika*, 1936, 1, 51-60.
3. Royer, Elmer B. A machine method for computing the biserial correlation coefficient in item validation. *Psychometrika*, 1941, **6**, 55-59.

Manuscript received 4/16/51.
Revised manuscript received 6/4/51.

PSYCHOMETRIKA—VOL. 17, NO. 1
MARCH, 1952

FACTOR ANALYSIS OF THE ARMY AIR FORCES SHEPPARD FIELD BATTERY OF EXPERIMENTAL APTITUDE TESTS*

J. P. GUILFORD
UNIVERSITY OF SOUTHERN CALIFORNIA
BENJAMIN FRUCHTER
UNIVERSITY OF TEXAS
AND
WAYNE S. ZIMMERMAN
BRANDEIS UNIVERSITY

A factor analysis was made of 39 experimental printed aptitude tests and seven reference tests selected from the Army Air Forces Aircrew Classification Battery. Thirteen factors were extracted and two independent orthogonal rotational solutions were completed. Twelve factors were interpreted. Of these, seven were clearly identifiable with previously known factors: numerical, nerceptual-speed, spatial-relations, visualization, visual-memory, paired-associates-memory, and length-estimation factors. A planning factor was not as clearly identifiable. A reasoning factor was probably a composite of two or more factors that failed to separate. A new factor possibly has to do with orientation with respect to the points of the compass. Two factors were doublets, each apparently specific to one kind of test. Better conceptions were gained of the spatial-relations and visualization factors and of the kinds of tests that measure them best. Efforts to improve measures of unique factors were not uniformly successful. The attempt to duplicate a psychomotor test rather directly by analogy in printed form failed almost completely.

Introduction

In the late stages of World War II, Psychological Units of the Army Air Forces Aviation Psychology Research Program had developed a large number of printed experimental tests for which no validity or factor-analysis information had been obtained. In the

*For support of a large part of the investigation the senior author is indebted to the Social Sciences Research Council for a grant-in-aid. Dr. Fruchter supervised the computational work in the extraction of factors and in one of the rotational solutions. Much of the work was done while he was employed by the Air Training Command Human Resources Research Center. While the opinions and conclusions are those of the authors, they wish to express their appreciation to that organization and especially to Dr. John T. Dailey for making the facilities possible and to Mr. William B. Lecznar for technical and computational assistance.

construction of the tests, efforts had been made to achieve better and more unique measures of certain primary abilities, to achieve a better understanding of those abilities, and to determine whether other, hypothesized, factors would be found. Since aircrew training had been very materially reduced in 1945, it was very unlikely that the trainees who were tested at the time would yield validation data. On the other hand, the validities of the better known factors had been so well estimated that it would be possible to make reasonable guesses of the minimum validities of the new experimental tests from a knowledge of their loadings in those factors (5, p. 843). All of these considerations made factor-analysis studies of the new tests extremely desirable.

Two large experimental testing projects were accordingly planned and executed during the spring and summer of 1945. In the first one, a group of 45 experimental tests was administered to a large sample of Aviation Students on the day immediately following the battery of 20 classification tests, at Sheppard Field. This set of tests became known as the Sheppard Field Battery. At that time, the Aviation Students were almost entirely at the eighteen- and nineteen-year levels. The total number involved in the study was 8,158, but since only one day's time was available for experimental testing, and since not all 45 tests could be completed in one day, not all of this sample took all tests. The 45 tests were subdivided into five half-day batteries of approximately nine tests each. Each sub-battery was administered in combination with every other sub-battery to approximately 400 students. Within sub-batteries the correlations were based upon nearly 1,600 students. The correlations between experimental and classification tests were based upon similar numbers. The classification-test intercorrelations were based upon the entire sample of 8,158 students. The entire correlation matrix for the 65 tests has been published in one of the Army Air Forces Reports (5, p. 902), with full particulars about sample N's. The portion of that matrix analyzed for this report is given in Table 1.

To be more specific concerning objectives for the experimental tests, special efforts had been made to measure better the factors of perceptual speed, spatial relations, visualization, visual memory, and length estimation. There were also efforts to clarify the nature of the factors of space, visualization, reasoning, and planning, particularly. It was thought quite possible that two new factors would be found, one involving space tests in which compass directions played a prominent role, and one involving tests of ability to resist illusions in geometric patterns. These objectives will be elaborated upon and others mentioned as the tests are described.

The analysis to be reported here did not include all 65 tests, for several reasons. The classification tests (September 1944 battery) had been analyzed previously and the results reported (5, p. 812). Certain classification tests were included in the present analysis because they would help to solve for some of the known factors that would be expected in the experimental data. By excluding others that had little relation to the experimental tests, the number of factors with which it was necessary to deal was reduced. The classification-battery factors of psychomotor coordination, judgment, mechanical knowledge, and verbal comprehension were thus ruled out. Two experimental tests of the last two factors were also eliminated when it was found that they correlated so little with other experimental tests. The two illusion tests were also omitted because it was likely that they would merely add a doublet factor to the structure. There remained, then, 46 tests in the analysis, seven of which were classification tests included for defining purposes. The 46 tests are described very briefly on the following pages.*

Description of the Test Variables

(1)† Map Memory, CI505BX1: This test was designed to measure pictorial memory. A schematic map is studied for four minutes, after which verbally stated printed questions must be answered regarding it.

(2) Figure Analogies, CI212AX1: This is a version of the well-known figure-analogies test.

(3) Spatial Visualization II, CI203A: The examinee reads a verbal description of a solid block of wood, its sides painted different colors, being cut into smaller blocks. His task is to visualize these cutting operations so that questions can be answered regarding the resulting number of blocks of a given size and color.

(4) Planning Air Maneuvers. CI408AX2: The problem is for the examinee to ascertain the quickest and most economical way to "sky write" certain letter pairs.

(6) Map Distance CP626B: On a schematic map a number of towns are located around a given reference point. The examinee's task is to estimate which one of any given pair of towns is closer to the ref-

*For more complete descriptions of these tests including sample items and statistical data see (5).

†The number preceding each test name and the code number following it are the designations used by the Air Forces (5, pp. 901-902).

erence point.

(7) Estimation of Length, CP631B: In the first section of this test the examinee must attempt to estimate which one of several given lines is exactly the same length as one of five standards. In the second part he tries to pick the one line that is exactly double the length of a standard.

(8) Speed of Identification, CP610C: Groups of five objects are shown, four of which in each group are identical with four objects shown in an adjacent group. The examinee's task is to designate matching pairs of objects.

(9) Memory for Plane Silhouettes, CI503AX1: The examinee studies silhouettes of top and side views of four to eight airplanes for eighty seconds, then takes a two-minute matching test on the scrambled views of the same airplanes.

(10) Directional Orientation, CP515F: Items consist of pairs of circular sections from aerial photographs. Although both pictures in a pair are identical, one is rotated. The examinee's task is to determine the compass direction of the rotated picture in relation to a compass direction marked on the unrotated picture.

(11) Visualization of Maneuvers, CI657CX2: A single view of an airplane is pictured in a starting position. A simple maneuver is described, such as a turn or a bank of a certain number of degrees. The examinee's task is to select the one of five alternative pictures that correctly portrays the airplane's position following the prescribed maneuver.

(12) Planning a Circuit, QP 901A-I: Each item presents an electrical-circuit diagram with intersecting and intermeshed wires and several sets of terminals. The task is to trace the circuits visually and to determine at which pair of terminals a battery should be placed in order to complete the circuit through a meter.

(13) Path Tracing, QP 901A-V: Items are similar to those on the McQuarrie Pursuit Test. In each of several diagrams, lines running in irregular fashion cross and recross one another. The examinee's task is to trace visually each line from its beginning and to mark its point of termination.

(14) Maze Tracing, QP 901A-VII: In a full-page, complicated printed maze, a number of points in the pathways are marked by letter. The items are pairs of letters. The examinee's task is to tell whether, in each item, the pathway between any two letters is clear or is blocked.

(15) Formation Visualization, CP814A: Each item shows in

silhouette, a top and side view of a formation consisting of either two or three airplanes. The examinee's problem is to select from five choices the one that portrays the formation from a front view.

(17) Visual Memory, CI514A: The examinee is given one minute to study a large aerial photograph. He then turns the page and selects from among several small photographs those that duplicate portions of the large one.

(18) Figure Classification, CI213AX1: The task is to select from five alternatives the geometric figure that has the characteristics common to the three stimulus figures that set the problem.

(19) Spatial Visualization I, CI204AX1: For each item there are two or three illustrations to show, step by step, how a sheet of paper is folded and then cut. The examinee's problem is to select the one of five alternative answers that correctly illustrates how the sheet will look when unfolded.

(20) Map Planning, CI412AX1: A diagramatic map is shown on which streets are represented, some of which are blocked by bomb damage. The examinee must plan quickly the shortest passable routes for military vehicles to travel through the damaged areas.

(21) Object Recognition, CP523A: This test is a revision of Thurstone's "Cubes." The examinee's task is to select the one of five cubes that portrays correctly a turned or rotated position of a given key cube.

(23) Position Orientation, CP526B: This is an adaptation of Thurstone's "Hands." In each item are shown five drawings of hands, arms, legs, eyes, or feet. The examinee's task is to determine quickly whether each drawing represents a right or a left member of the body.

(24) Aerial Orientation, CP520C: For each item the stimulus is a cockpit view of a shoreline. Adjacent to each stimulus picture are five photographs of an airplane in different attitudes. The examinee's problem is to match the cockpit view of the shoreline with the airplane position from which that view would be seen.

(25) Object Identification I, CP521A-I: This is a revision of Thurstone's "Flags." Silhouettes of planes, trucks, guns, tanks, and ships are presented. The examinee's task is to select from five alternative answers those rotated illustrations that show the same side of an object as that shown in a key illustration.

(26) Object Identification II, CP521A-II: This form is similar to variable (25) except that flags are presented instead of planes, trucks, etc.

(27) Plane Position Memory, CI512A: On each of several study pages nine airplanes are shown. Following a two-minute study period the same nine airplanes located differently on a page are shown to the examinee. His problem is to recall the row in which each airplane appeared on the study and the direction in which it was headed.

(28) Decoding, CI214AX2: The test problems require the examinee to match groups of short words written in a code of signal flags with the same words written in the English alphabet.

(29) Route Planning, CI411AX1: The examinee must plan paths successively from four points on the periphery of printed mazes to goal boxes in their centers.

(30) Flight Formation, CI654AX5: The examinee must select from among five alternatives the position of three airplanes after certain moves from a given formation have been described verbally.

(31) Aerial Landmarks, CP525C: Paired aerial photographs are presented, one of which shows a vertical view and the other an oblique view of the same terrain. The examinee's task is to find on the oblique view points given on the vertical view.

(32) Pattern Assembly, CP804A: This is a paper-form-board type of test. Each item requires selection of the one of five assembled patterns that correctly represents how the unassembled parts, illustrated in a separate panel, would look when fitted together.

(33) Block Counting, CP512B-4: In various piles of blocks the examinee's task is to count as rapidly as possible the number of adjacent blocks that are touching the sides of certain designated ones.

(34) Discrimination Reaction Time (paper), CP634A-I & II: This test was designed to duplicate factorially in printed form the psychomotor test of the same name. The examinee responds to a pair of black-and-white dots by marking on a specially designed answer sheet in one of four directions—up, down, left, or right. Responses are determined by the relative positions of the dots.

(35) Discrimination Reaction Time (paper), CP634A-III & IV: Parts III and IV call for the same type of responses as in Parts I and II, but the direction of marking is determined by the arrangement of three circles, one white, one black, and one with a cross.

(36) Plane Name Memory, CI503AX2: In each of two parts, the examinee studies for four minutes a group of 20 plane silhouettes, the name of each appearing below it. Following the study period the silhouettes are shown on another page with five names listed under each. The examinee selects the name he thinks appeared originally with each silhouette.

(37) Planning a Course, CI406AX3: The Problem in this test consists of tracing lines through a diagram like that of city streets. The directions for the tracings are determined by a learned signal code. Signals are varied, however, throughout the maze, necessitating changes in the mode of response as the maze is traversed.

(38) Compass Orientation, CI660A: In each item one of the four compass directions, north, south, east, or west, is presented verbally as the initial flight direction of an airplane. Then a turn, either left or right, is specified. The examinee's task is to record the new compass direction of flight.

(39) Competitive Planning, CI409AX2: This test is based on the familiar completion-of-squares game, sometimes called "Squares" or "Boxes." In the test, the examinee must plan moves for two opponents according to given rules, so that each completes as many squares as possible in a rectangular diagram, and must indicate how many squares are completed by each opponent.

(40) Camouflaged Outlines, CP821A: This is a variation of Gottschaldt's Figures test. The examinee's task is to detect rapidly simple outline figures within complex designs.

(41) Angle Estimation, CP218A: The test is composed of photographs of military vehicles taken from the air at ten-degree-angle intervals ranging from zero to ninety. The examinee's task is to judge the angle at which each photograph was taken.

(42) Spatial Reasoning, CI211BX2: This is a revision of Thurstone's "Marks." It requires the examinee to detect principles governing the placement of letter symbols in spatial patterns of dashes and gaps.

(51) Instrument Comprehension, CI616C*: Each item shows two airplane instruments, a compass and an artificial horizon, followed by photographs of an airplane in five different attitudes. The examinee must select the photograph showing the airplane in an attitude agreeing with the instruments' readings.

(52) Mechanical Principles, CI903B: This test is composed of items similar to those in the familiar Bennett and Fry "Mechanical Comprehension Test." For each item the examinee must select from alternate answers the one that describes most accurately what is happening or will happen in a pictured situation.

(53) Speed of Identification, CP610A: This test is similar to test (8) in this series except that the items are composed of airplane silhouettes, the perceptually distinguishable differences are not as

*This test and those following were in the Aircrew Classification Battery.

gross, and in most paired views one is rotated.

(54) Numerical Operations I, CI701B: Items are composed of simple addition and multiplication problems of the true-false type, which are printed on an IBM answer sheet.

(55) Numerical Operations II, CI701B: This part, a continuation of the preceding test, presents simple division and subtraction problems with five-choice answers.

(59) Arithmetic Reasoning, CI206C: This test is composed of arithmetical problems pitched at a level difficult enough to depend upon reasoning in the solutions.

(63) Complex Coordination, CM701A: This is a psychomotor test in which patterns of lights are presented whose positions can be matched by making stick-and-rudder manipulations. Proper adjustments of the stick and rudder cause new light patterns to be presented automatically. The number of new patterns elicited within a given time interval is the examinee's score

Analysis of the Data

The factor-analysis procedure. Factors were extracted from the 46-by-46 correlation matrix by a combination of the multiple-group and the complete centroid methods (8).The multiple-group method was considered applicable for the extraction of the first seven factors (numerical, perceptual speed, spatial relations, general reasoning, space II, visualization, and visual memory). Residuals were computed and the remaining factors were extracted by the centroid method.* Extractions were continued until the product of the two highest centroid loadings (.049) was not greater than the standard error of the original correlation of the same two variables (.047). Six centroid factors were extracted, which made a total of thirteen factors in all. Positions of the variables were then plotted and pair-by-pair orthogonal rotations were made using Zimmerman's simplified graphical method (9).

Factors. extr.

Rot.

Independent rotational solutions were completed by two individuals. Criteria guiding the rotations included simple structure, positive manifold, psychological meaningfulness, and concordance with previous well-established factor-analysis findings. Tables 1, 2, 3, and 4 pre-

*The application of the multiple-group method included a rotation to an orthogonal reference frame. It was assumed that the spaces of the factors extracted by the two methods were mutually orthogonal. One indication that this is so is given by the fact that communalities before and after rotation of the two systems together agreed very well.

sent the correlation matrix, the unrotated factor loadings, and the two sets of rotated factor loadings, respectively.

Description of the factors. In the first of the two solutions all thirteen factors were rotated. Centroid axis XIII, however, failed to yield a meaningful result. In the second solution, centroid axis XIII was discarded prior to rotation. No attempt was made to reconcile the two solutions. Although there are noteworthy differences between the two solutions, the twelve rotated factors are all sufficiently alike to warrant the same identifications. The most glaring difference is on the visual-memory factor (rotated factor IX) in which the leading test in the second solution occupies a relatively insignificant position in the first. Results from the two sets of rotations are described and compared in the factor descriptions that follow. Tests with loadings of .30 or higher in either solution are listed under each factor.

Rotated Factor I, Numerical. The three tests, Numerical Operations II, Numerical Operations I, and Arithmetic Reasoning are weighted most heavily. No other primarily numerical tests were included among the forty-six variables analyzed.

		Loadings	
Test No.	Test Name	I	II
55	Numerical Operations II	.80	.73
54	Numerical Operations I	.77	.70
59	Arithmetic Reasoning	.46	.51
8	Speed of Identification C	.36	.17
42	Spatial Reasoning	.34	.32
28	Decoding	.33	.39
23	Position Orientation	.32	.24
38	Compass Orientation	.32	.18

Rotated Factor II, Perceptual Speed. The C form (experimental) of Speed of Identification is more heavily weighted than the A form (classification), as had been predicted (5). It is interesting to note

		Loadings	
Test No.	Test Name	I	II
8	Speed of Identification C	.58	.57
53	Speed of Identification A	.53	.50
32	Pattern Assembly	.46.	.42
33	Block Counting	.38	.29
7	Estimation of Length	.33	.32
6	Map Distance	.32	.41

that the simpler C form, despite its greater perceptual-speed loading, is more complex factorially, carrying in addition significant weights in both length-estimation and numerical factors. The loading for Speed of Identification A is notably less than had usually been found (.64) but is reasonably close to that carried in the Air Force analyses of the September 1944 Battery (.58).* Pattern Assembly appears with an unusually heavy perceptual-speed weight. Its perceptual content was evident in previous Air Force analyses but not so prominently (.26). In the first rotational solution the loading for Block Counting is in line with its previous Air Force weight of .43. In previous Air Force analyses, Map Distance was formerly unweighted in the perceptual-speed factor. Its presence here could suggest a possible correlation of the factor with that of length estimation.

Rotated Factor III, Spatial Relations. As may be seen in the list of tests and loadings below, Aerial Orientation leads the other tests, with Visualization of Maneuvers C, running a close second. Instrument Comprehension, ranking third in one solution and fourth in the other, and Complex Coordination, fifth in one and seventh in the other, were the best measures of Spatial Relations found in previous Air Force analyses. Aerial Orientation was developed in an attempt to measure the same space factor with even greater strength and purity, apparently with some success. Visualization of Maneuvers C is a less difficult and consequently more speeded version of the original Form A of the test. It was expected, therefore, to be a better measure of space than of the more intellectually difficult visualization. This proved to be the case. The fact that Instrument Comprehension and Complex Coordination rank somewhat below their previous standings on the space factor suggests that the direction of the space axis in this analy-

		Loadings	
Test No.	Test Name	I	II
24	Aerial Orientation	.61	.62
11	Visualization of Maneuvers C	.59	.58
15	Formation Visualization	.44	.44
51	Instrument Comprehension	.41	.47
9	Memory for Plane Silhouettes	.35	.26
31	Aerial Landmarks	.34	.21
63	Complex Coordination	.30	.40
52	Mechanical Principles	.27	.36
41	Angle Estimation	.24	.30

*An analysis of the intercorrelations of the September 1944 Classification Battery tests contained in the Sheppard Field Matrix (5).

sis may be altered somewhat. In the second rotational solution the values for these tests approach their previous loadings more closely. The next most significantly weighted test, Formation Visualization, devised as a measure of visualization primarily, contributed variance almost equally to this factor and the visualization factor described below.

Hypotheses regarding the nature of the spatial-relations and the visualization factors have been advanced by several investigators (1, 2, 3, 4, 6, 7, 10, 12). With Aerial Orientation and Visualization of Maneuvers C leading all other tests by a substantial margin, emphasis for this space factor seems to be placed upon empathic involvement and directional discrimination. The examinee must "place himself" in the cockpit of the airplane and quickly determine the direction of motion involved—right or left, up or down— depending upon a correct appraisal or "feeling" for the stimulus arrangement. Orientation is with respect to his own body. As might be expected this ability proved to be one of the most prominent in the pilot-training criterion.

Rotated Factor IV, Visualization. A separation between the visualization and spatial-relations factors occurred first in Air Force analyses (11). The tests Mechanical Principles and Spatial Visualization I represented the visualization factor best in those analyses

		Loadings	
Test No.	Test Name	I	II
3	Spatial Visualization II	.63	.60
19	Spatial Visualization I	.61	.60
52	Mechanical Principles	.60	.55
41	Angle Estimation	.50	.45
15	Formation Visualization	.45	.40
11	Visualization of Maneuvers C	.44	.26
21	Object Recognition	.41	.37
2	Figure Analogies	.37	.41
51	Instrument Comprehension	.37	.27
31	Aerial Landmarks	.34	.28
10	Directional Orientation	.34	.30
59	Arithmetic Reasoning	.31	.33
26	Object Identification II	.31	.23
13	Path Tracing	.30	.34
40	Camouflaged Outlines	.30	.34
29	Route Planning	.25	.33
12	Planning a Circuit	.22	.32
14	Maze Tracing	.20	.33

with Spatial Visualization II holding a lesser but still prominent position. As may be seen in the list above, these three tests clustered together, leading all others in weights in this factor. The fourth most heavily loaded test, Angle Estimation, had once been considered as a potential representative of a new factor. It is interesting to note, therefore, its heavy weight in visualization.

The leading visualization tests seem to involve a "mental" manipulation of objects in space. It is usually necessary to move, turn, twist, or rotate an object or objects in imagination and to recognize a new appearance, position, or condition after prescribed manipulations. This visualization factor probably corresponds more closely than does that of spatial relations to Kelley's and Thurstone's space factor.

Rotated Factor V, Reasoning. In previous Air Force analyses as many as three different reasoning factors have been described. The factor represented by the tests listed below appears to be a composite of at least two of these factors.

		Loadings		AAF Factors for Reasoning		
Test No.	Test Name	I	II	I	II	III
18	Figure Classification	.60	.51	.03	.16	.32
2	Figure Analogies	.48	.36	.34	.40	.31
59	Arithmetic Reasoning	.46	.35	.47	—	—
30	Flight Formation	.45	.59	.16	—	—
37	Planning a Course	.42	.34	.24	—	—
3	Spatial Visualization II	.41	.26	.39	—	—
42	Spatial Reasoning	.38	.36	.45	.05	.38
21	Object Recognition	.34	.27	—	—	—
28	Decoding	.31	.25	.36	.30	.37

Mathematics tests similar to Arithmetic Reasoning headed the list of the Air Force reasoning-I factor tests. Figure Analogies was one of the best tests of reasoning II, and Spatial Reasoning and Decoding were among the best tests of reasoning III. The first rotational solution is more congruent with the Air Force's general-reasoning factor (reasoning I).

Rotated Factor VI, Paired-Associates Memory. This cluster of tests obviously represents memory of some kind. In previous Aiı Force analyses Plane Name Memory was loaded heavily in two memory factors, one undefined and the other called paired-associates memory. Memory for Plane Silhouettes, however, was weighted only in paired-associates memory, which probably corresponds with Thurstone's factor M, or rote memory.

Test No.	Test Name	Loadings I	 II
9	Memory for Plane Silhouettes	.59	.59
27	Plane Position Memory	.58	.59
36	Plane Name Memory	.48	.47

Rotated Factor VII, Object Identification (doublet). In one analysis of AAF perceptual tests, a second space factor was distinguished from the better know spatial-relations factor (5). The two tests loaded significantly on "Space II" were Thurstone's Hands, and Flags, Figures, and Cards. Since Object Identification I and II are variations of Flags, and Position Orientation is an adaptation of Hands, it was assumed that these three tests might define a second space factor in this analysis. But it can be seen in the following table that only Parts I and II of Object Identification have high loadings which make this factor a doublet, specific to this test. The factor space II was thus not verified by the AAF version of Thurstone's Hands test.

Test No.	Test Name	Loadings I	 II
25	Object Identification I	.64	.62
26	Object Identification II	.55	.57
42	Spatial Reasoning	.30	.22
23	Position Orientation	.29	.32

Rotated Factor VIII, Planning Speed. With the exception of Maze Tracing and Block Counting all of the tests in the list below had been analyzed previously in Air Force studies. In those studies, factors appeared involving several tests in the list. Planning Air Maneuvers, Planning a Course, Route Planning, and Spatial Reasoning, in at least one analysis, received loadings of .33 and above on a factor called integration III. Integration III was described very tentatively as the ability to keep in mind and integrate a number of de-

Test No.	Test Name	Loadings I	 II
12	Planning a Circuit	.53	.43
14	Maze Tracing	.50	.57
42	Spatial Reasoning	.34	.31
20	Map Planning	.32	.38
13	Path Tracing	.29	.29
33	Block Counting	.26	.38
29	Route Planning	.26	.38
4	Planning Air Maneuvers	.23	.28
37	Planning a Course	.23	.28

tailed instructions. Planning Air Maneuvers and Planning a Circuit showed substantial loadings along with a moderately loaded Map Planning on a factor labeled planning. The factor was not satisfactorily interpreted since other planning tests in the same analysis did not show significant weight. Owing to the inclusion of a larger number of planning tests in this battery, rotated factor VIII is probably more stable than either of the two factors mentioned. The addition of Maze Tracing to the analysis seems to have brought out the variance that it holds in common with Planning a Circuit. The emphasis has shifted away from Planning Air Maneuvers, which led in relation to the AAF's planning factor. Since Planning Air Maneuvers involves more complex and difficult problems than either Planning a Circuit or Maze Tracing, speed is emphasized in the new factor. Thus the term "planning speed" is suggested as a title. The ability represented is obviously something more complex than speed in visual tracing of lines or paths since Path Tracing, the AAF's counterpart of the Pursuit test, is only moderately weighted.

Rotated Factor IX, Visual Memory. A visual-memory factor was anticipated because of the inclusion in the matrix of Map Memory, which represented the factor best in Air Force analyses, along with

Test No.	Test Name	Loadings I	 II
1	Map Memory	.50	.43
17	Visual Memory	.36	.43
2	Figure Analogies	.29	.37
3	Spatial Visualization II	.26	.34
27	Plane Position Memory	.25	.32
31	Aerial Landmarks	.19	.50
15	Formation Visualization	.04	.33
11	Visualization of Maneuvers	.03	.37

the new test Visual Memory, which was constructed for the special purpose of measuring the factor that bears its name.

This is the only factor in which the results differ greatly in the two rotational solutions. In the first solution, Map Memory holds the leading position with Aerial Landmarks ranked far down the list. In the second, Aerial Landmarks defines the factor best, with Map Memory not far behind.

The reader may well question why a disparity of this magnitude should exist between the two loadings for Aerial Landmarks. The answer probably lies in the difference in weight given by the two analysts to the various guiding criteria. In solution I, probably greater weight was given to psychological meaningfulness and invariance when rotations were more or less indeterminant with respect to simple structure or positive manifold. An examination of the content of Aerial Landmarks would lead one to expect significant weights in such factors as perceptual speed, visualization, and spatial relations. It would seem logical that matching points on the two photographs should tap somewhat the same abilities measured by Spatial Orientation II, which also presents sections of aerial photographs to be matched.* Spatial Orientation II, an almost pure measure of perceptual speed, contains items pitched at a difficulty level somewhat below that of Aerial Landmarks. The latter test adds to the perceptual element the complication of rotation and right-left determination, features believed to involve visualization and space, respectively. It is difficult to think of visual memory playing a dominant role in solving Aerial Landmarks items. Visual memory is defined as the ability to retain an impression of pictorial material (as if photographically) and to recognize it after a short time interval (6). The vertical and oblique aerial views in Aerial Landmarks involve two different retinal pictures.

Rotated Factor X, Length Estimation. In several Air Force analyses Pattern Assembly proved to be the best representative of a fac-

		Loadings	
Test No.	Test Name	I	II
6	Map Distance	.56	.43
7	Estimation of Length	.39	.36
33	Block Counting	.32	.30
8	Speed of Identification C	.31	.26
32	Pattern Assembly	.28	.31

*See (5) for a description of Spatial Orientation II.

tor labeled length estimation. The test Estimation of Length was constructed in an attempt to improve the measurement of this factor. As may be seen, Map Distance leads all other tests with Estimation of Length next in line and Pattern Assembly holding a position of lesser importance.

It is interesting to note length-estimation variance in problems involving estimations of the number of pieces contained within a given dimension (as in Block Counting), and in estimation of comparative sizes of perceptual patterns (as in Speed of Identification C and Pattern Assembly). It is easier to rationalize strong length-estimation content in Map Distance and Estimation of Length than it is in Pattern Assembly, and to that extent the present findings make more "psychological sense" than do the AAF findings.

Rotated Factor XI, Discrimination-reaction-time (*doublet*). The two forms of the test heading the list below were constructed in an effort to measure with a printed test the factors in the psychomotor test of the same name. The latter was characterized factorially by the following factors and loadings: spatial relations, .42; psychomotor precision, .35; perceptual speed, .22; and visualization, .20 (5). The printed tests failed almost completely to measure the factors of spatial relations, perceptual speed, and visualization, since the loadings in these factors were very small.

Since there are no other known tests of psychomotor precision in this matrix, it is possible that the two forms of Discrimination Reaction Time have achieved the goal of measuring this factor. This hypothesis is very unlikely, however, since previous results have shown that there is some psychomotor-precision variance in other tests in the present battery and since this factor did not emerge in them in the present analysis. In developing the printed discrimination-reaction-time tests, it was believed that the chief variance that would carry over from the psychomotor form of the same test would

		Loadings	
Test No.	Test Name	I	II
34	Discrimination Reaction Time I & II	.50	.50
35	Discrimination Reaction Time III & IV	.46	.44
29	Route Planning	.37	.27
27	Plane Position Memory	.30	.24
20	Map Planning	.30	.23
14	Maze Tracing	.30	.05
38	Compass Orientation	.09	.38

be in the spatial-relations factor. This was based upon one hypothesis that this space factor is essentially a matter of decision as to direction of movement. The absence of spatial-relation variance in the printed forms is evidence against that hypothesis. In this connection, attention is called to the fact that the space factor is well measured by printed tests of a different type.

Rotated Factor XII, Compass Orientation. Before the analysis there was speculation as to whether tests that involved the use of points of the compass would bring out another space factor or whether their variances could be largely accounted for in terms of the well-established spatial-relations factor. The results are not very decisive concerning this question. With only a single substantially weighted test the factor must be explained, if at all, by the content of the test itself. Without supporting tests or previous knowledge of Compass Orientation, common variance cannot be claimed. It would be very interesting, however, to find that the appreciation of spatial arrangements is mediated by two abilities, one with reference to the body of the observer (spatial relations) and the other with reference to compass points. The two frames of reference might be expected to yield separate abilities.

		Loadings	
Test No.	Test Name	I	II
38	Compass Orientation	.66	.57
41	Angle Estimation	.32	.08
20	Map Planning	.23	.32

Rotated Factor XIII, Residual. After rotation Block Counting gained a loading of .32 on Axis XIII in the first solution, but no other test approached even this low figure. Although a residual ideally should appear with diminished loadings only, it is very doubtful that a significant factor is represented.

Conclusions

1. In general, previously obtained factors, in the AAF results and elsewhere, were confirmed by this study, as were their relationships to specific tests.

2. Better tests, in terms of increased factor loadings, seem to have been developed for the factors of perceptual speed and spatial relations. Two previously developed experimental visualization tests were found to have higher loadings in that factor than had formerly been supposed.

3. Tests designed as improved measures of the factors of length estimation, space II, and visual memory seem to be inferior to previous ones for the same factors.

4. Certain hypotheses concerning the nature of the spatial-relations and visualization factors were supported: (1) that the AAF spatial-relations factor is an ability to perceive relations of objects in space with respect to the observer's body, an orientation in which the human body is the frame of reference; and (2) that visualization is the ability to manipulate visual objects mentally.

5. The usual reasoning factors failed to separate, probably because of an insufficient number and variety of definitive reasoning tests in the battery.

6. The effort to reproduce the factorial composition of a psychomotor test (Discrimination Reaction Time) in printed forms failed almost completely. The measurement of most of the factors involved, however, has been achieved in various other printed tests.

7. There is some indication of a new space factor in which orientation depends upon the compass points as a frame of reference. This hypothesis is worth serious investigation.

REFERENCES

1. Comery, A. L. A factorial study of achievement in West Point courses. *Educ. psychol. Meas.*, 1949, 9, 193-209.

2. Dudek, F. The dependence of factorial composition of aptitude tests upon population differences among pilot trainees. *Educ. psychol. Meas.*, 1948, 8, 613-633; 1949, 9, 95-104.

3. Fruchter, B. The nature of verbal fluency. *Educ. psychol. Meas.*, 1948, 8, 33-47.

4. Fruchter, B. Factorial content of right-response and wrong-response scores in a battery of experimental aptitude tests. Unpublished Ph.D. dissertation, University of Southern California, 1948.

5. Guilford, J. P. (Ed.) Printed classification tests. Army Air Forces Aviation Psychology Research Program, Report No. 5, Washington: U. S. Government Printing Office, 1947.

6. Guilford, J. P., and Zimmerman, W. S. Some AAF findings concerning aptitude factors. *Occupations*, 1947, 26, 154-159.

7. Michael, W. B. Factor analyses of tests and criteria: a comparative study of two AAF pilot populations. *Psychol. Monog.*, 1949, 63, No. 298.

8. Thurstone, L. L. Multiple-factor analysis. Chicago: Univ. of Chicago Press, 1947.

9. Zimmerman, W. S. A simple graphical method for orthogonal rotation of axes. *Psychometrika*, 1946, 11, 51-55.

10. Zimmerman, W. S. Isolation, definition, and measurement of spatial and

visualizing abilities. Unpublished Ph.D. dissertation, University of Southern California, 1948.

11. Zimmerman, W. S. Visualization. Chapter 12 in J. P. Guilford, (Ed.) Printed Classification Tests. Washington: U. S. Government Printing Office, 1947.
12. Zimmerman, W. S., and Howe, J. A., Jr. Spatial Tests. Chapter 19 in J. P. Guilford, (Ed.), Printed Classification Tests. Washington: U. S. Government Printing Office, 1947.

Manuscript received 10/16/50.
Revised manuscript received 3/17/51.

TABLE I
Intercorelation Matrix*

Test No.	Test Title	1	2	3	4	6	7	8	9	10	11	12	13	14	15	17	18	19	20	21	23	24	25	26
1	Map Memory		39	35	25	25	20	29	28	31	29	29	23	33	29	36	26	39	36	32	22	39	31	34
2	Figure Analogies			59	34	25	22	33	27	42	47	40	30	46	48	31	47	55	35	47	21	42	35	33
3	Spatial Visualization II ✓				38	23	22	26	34	42	51	40	38	41	54	25	41	63	34	50	19	42	37	42
4	Planning Air Maneuvers					13	13	16	24	28	35	29	26	33	38	25	17	33	28	29	15	27	25	19
6	Map Distance						40	46	14	29	19	22	26	27	15	26	24	29	29	27	27	25	27	26
7	Estimation of Length							39	09	23	10	15	17	21	09	10	12	14	23	14	24	13	19	23
8	Speed of Identification								24	37	26	33	30	39	23	28	29	23	38	21	31	17	25	24
9	Memory for Plane Silhouettes									31	43	33	24	25	38	31	15	45	21	26	21	45	29	29
10	Directional Orientation										48	41	37	38	43	33	23	47	28	42	29	41	41	40
11	Visualization of Maneuvers											38	34	30	57	28	27	48	23	48	31	60	42	38
12	Planning a Circuit												39	48	44	28	28	47	31	31	23	42	33	38
13	Path Tracing													33	32	16	16	38	31	25	23	29	25	34
14	Maze Tracing														41	29	23	43	44	34	29	32	31	33
15	Formation Visualization															29	30	57	30	49	32	57	39	38
17	Visual Memory																17	31	23	22	16	24	23	22
18	Figure Classification ✓																	33	17	32	14	27	23	25
19	Spatial Visualization I ✓																		33	47	26	43	38	41
20	Map Planning																			28	28	25	32	33
21	Object Recognition																				30	38	39	40
23	Position Orientation																					30	38	41
24	Aerial Orientation																						32	32
25	Object Identification I																							62
26	Object Identification II																							
27	Plane Position Memory																							
28	Decoding																							
29	Route Planning																							
30	Flight Formation																							
31	Aerial Landmarks																							
32	Pattern Assembly																							
33	Block Counting																							
34	Discrimination Reaction Time I & II																							
35	Discrimination Reaction Time III & IV																							
36	Plane Name Memory																							
37	Planning a Course																							
38	Compass Orientation																							
39	Competitive Planning																							
40	Camouflaged Outlines																							
41	Angle Estimation																							
42	Spatial Reasoning																							
51	Instrument Comprehension																							
52	Mechanical Principles																							
53	Speed of Identification																							
54	Numerical Operations I																							
55	Numerical Operations II																							
59	Arithmetic Reasoning																							
63	Complex Coordination																							

*Decimal points omitted.

TABLE 1 (Continued)
Intercorelation Matrix*

Test

TABLE 1 (Continued)
Intercorelation Matrix*

Test No.	Test Title	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	51	52	53	54	55	59	63
1	Map Memory	40	36	26	30	35	25	38	22	23	36	34	35	23	22	20	32	28	18	26	25	34	50	31
2	Figure Analogies	28	46	37	45	41	36	49	19	24	31	48	36	34	41	22	48	31	38	27	19	26	52	30
3	Spatial Visualization II	21	44	41	32	42	32	52	18	20	27	39	30	33	42	35	39	34	48	14	07	13	26	20
4	Planning Air Maneuvers	13	25	28	24	29	19	34	08	10	07	27	09	23	21	18	28	22	27	32	26	26	19	22
6	Map Distance	15	30	23	20	11	36	42	24	23	20	17	24	14	29	–01	26	23	08	28	26	26	18	20
7	Estimation of Length	16	27	21	23	13	33	38	24	27	19	27	23	19	30	–02	25	20	05	42	39	38	24	32
8	Speed of Identification	17	43	25	35	19	33	47	32	38	26	35	43	20	35	12	39	21	02	37	04	08	15	28
9	Memory for Plane Silhouettes	50	30	30	23	39	28	32	18	17	34	17	06	16	28	21	19	33	28	31	24	26	36	32
10	Directional Orientation	28	48	39	32	45	29	41	32	30	14	39	14	21	39	28	40	36	37	30	14	18	34	36
11	Visualization of Maneuvers	34	37	34	23	52	24	42	23	27	14	35	11	20	43	42	36	46	50	26	19	22	28	21
12	Planning a Circuit	25	36	43	37	28	18	45	19	30	15	33	–07	21	30	22	36	31	36	26	08	14	26	36
13	Path Tracing	21	31	29	25	28	24	37	27	25	18	18	08	14	33	19	25	22	29	21	12	13	22	33
14	Maze Tracing	24	41	42	29	29	22	45	22	23	09	37	00	31	36	21	39	25	29	23	13	16	26	28
15	Formation Visualization	30	39	34	30	49	23	47	20	25	13	33	15	24	36	34	38	43	49	29	10	17	35	34
17	Visual Memory	32	23	17	21	35	21	31	16	21	22	22	15	19	26	19	28	22	15	23	13	13	18	20
18	Figure Classification	16	28	17	36	21	23	23	14	11	23	30	25	25	26	25	33	27	25	18	17	21	35	23
19	Spatial Visualization I	19	40	35	31	46	31	46	15	19	32	28	28	31	49	44	37	43	53	33	12	20	39	35
20	Map Planning	21	36	38	26	24	24	45	30	31	18	35	38	34	28	08	43	23	17	26	23	28	23	28
21	Object Recognition	16	41	29	32	34	19	34	19	20	28	34	38	33	33	31	37	33	39	20	22	29	40	29
23	Position Orientation	25	29	21	26	32	22	41	26	32	17	20	44	18	33	15	30	30	17	22	31	32	22	32
24	Aerial Orientation	23	26	28	27	41	17	39	16	23	18	29	31	27	26	39	33	51	42	26	11	16	28	40
25	Object Identification I	25	33	25	22	35	26	37	20	27	20	27	33	29	39	20	39	31	29	25	25	25	26	33
26	Object Identification II	15	32	27	26	32	31	38	25	23	11	32	35	28	35	18	40	31	30	27	29	29	28	33
27	Plane Position Memory		28	25	25	30	23	27	24	21	40	17	30	13	19	23	15	27	19	24	14	15	16	26
28	Decoding			36	41	34	22	39	31	33	20	32	32	27	31	16	37	27	16	30	33	37	40	21
29	Route Planning				36	36	21	43	27	27	10	30	27	27	32	31	30	25	31	21	12	17	27	30
30	Flight Formation					27	20	36	29	33	23	44	39	25	34	15	51	28	18	22	27	30	38	23
31	Aerial Landmarks						22	37	20	20	21	20	27	19	26	39	24	34	34	32	08	16	29	28
32	Pattern Assembly							39	19	17	13	13	19	11	30	22	17	16	16	35	12	13	13	25
33	Block Counting								31	35	17	36	37	26	40	31	44	35	30	37	25	28	31	39
34	Discrimination Reaction Time I & II									48	28	21	32	21	30	14	30	19	08	24	27	26	20	25
35	Discrimination Reaction Time III & IV										15	35	34	26	25	12	32	21	09	22	26	27	23	25
36	Plane Name Memory											17	26	13	25	15	19	22	07	23	16	15	17	11
37	Planning a Course												30	29	29	12	42	26	18	18	31	37	37	24
38	Compass Orientation													11	17	07	33	28	11	25	34	35	31	19
39	Competitive Planning														23	16	29	21	16	17	23	28	33	21
40	Camouflaged Outlines															24	36	25	24	27	19	20	28	26
41	Angle Estimation																15	31	36	21	02	06	17	26
42	Spatial Reasoning																	25	21	17	38	41	46	23
51	Instrument Comprehension																	39	28	28	17	22	30	36
52	Mechanical Principles																			17	–02	08	35	32
53	Speed of Identification																				15	17	12	28
54	Numerical Operations I																					67	40	16
55	Numerical Operations II																						50	12
59	Arithmetic Reasoning																							18
63	Complex Coordination																							

*Decimal points omitted.

46. Test

TABLE 2
Multiple-Group (I through VII) and Centroid Loadings (VIII through XIII)*

Test No.	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	h^2
1.	26	33	28	03	20	33	12	04	05	16	—10	—15	—17	51
2.	37	29	32	39	30	10	—06	05	05	18	02	07	—05	62
3.	38	36	34	55	11	05	00	07	—02	21	04	08	—08	77
4.	12	18	26	23	20	—03	01	19	11	10	—08	—04	06	28
6.	32	40	05	—03	04	—02	01	10	—25	07	—29	18	—14	48
7.	32	34	—04	—09	14	—01	01	15	—23	09	—09	08	—04	35
8.	48	51	—15	—08	27	—08	—05	20	—12	—05	—07	07	—04	67
9.	08	58	29	06	—02	29	08	—10	23	—28	—08	07	09	67
10.	31	39	31	24	10	—04	12	22	13	—08	07	11	—08	52
11.	20	41	58	14	02	—05	14	17	19	05	16	24	10	75
12.	14	37	35	25	24	—04	18	15	12	—23	—12	—08	—16	56
13.	16	28	21	27	08	07	09	28	—10	—15	—10	12	—05	38
14.	18	30	24	28	29	—02	08	40	08	—10	—14	—12	—18	59
15.	17	40	54	27	09	—09	09	09	17	—05	10	04	02	64
17.	20	38	09	08	16	22	04	10	24	08	—12	—12	08	38
18.	16	23	26	16	40	—01	03	—20	—08	14	08	22	—16	48
19.	20	46	38	48	08	06	03	—02	—02	04	—08	—04	—04	65
20.	20	35	13	11	33	05	17	21	—18	—05	—03	—21	—15	47
21.	32	21	36	34	12	07	14	—02	—03	10	13	07	—08	47
23.	39	22	24	—03	—01	05	28	10	—20	—12	03	—06	18	44
24.	17	36	67	—04	08	—03	05	05	07	07	02	03	—02	63
25.	31	29	25	18	07	08	58	—04	—04	—02	—05	05	05	64
26.	36	31	22	23	03	05	50	—05	—09	03	—09	—06	—05	61
27.	18	32	17	—04	12	56	01	14	13	—15	12	06	22	62
28.	44	33	10	15	26	05	04	11	08	—08	15	07	—19	49
29.	18	27	22	25	24	—01	00	30	—05	—05	13	—14	07	42
30.	35	23	20	07	52	—02	—05	—07	—04	—13	13	—07	05	55
31.	15	46	31	19	08	09	10	04	19	05	24	—10	11	50
32.	16	50	—08	17	14	00	08	08	—08	22	—05	18	16	46
33.	33	48	22	16	25	—06	03	31	—14	10	—08	—09	15	64
34.	33	27	02	—06	18	14	06	22	—24	—20	23	04	—03	44
35.	33	24	10	—06	23	—01	08	28	—16	—15	22	—05	—07	42
36.	19	30	09	06	16	44	—09	—17	—14	—08	—05	12	—10	44
37.	42	15	23	05	41	—04	04	06	08	05	04	—05	—15	47
38.	43	26	20	—07	23	14	17	—27	—36	12	27	—21	07	73
39.	32	16	20	11	13	—03	10	03	—05	04	04	—09	—23	28
40.	24	35	16	31	22	—02	11	08	—17	—11	—05	16	08	44
41.	05	32	38	24	—15	08	—06	06	—11	10	22	—15	17	46
42.	49	11	26	17	42	—07	16	05	—03	—05	—10	—03	02	57
51.	24	36	49	06	—08	03	00	—05	—06	—06	04	—02	—04	46
52.	04	24	52	41	—07	06	03	02	02	05	08	05	03	52
53.	20	63	00	00	00	00	02	01	00	01	00	00	—01	44
54.	80	—02	—03	—05	00	02	06	02	—08	—05	—03	05	—02	65
55.	81	02	03	05	00	—02	—04	—03	—03	—03	06	—06	—10	69
59.	56	01	29	34	17	00	—12	—03	04	08	10	06	—08	58
63.	17	38	35	04	02	—07	14	14	—14	—08	—02	05	03	38

*Decimal points omitted.

TABLE 3
Rotated (Orthogonal) Factor Loadings (Solution I)*

Test No.	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	h^2
1.	08	11	20	06	14	16	13	18	50	24	04	19	02	51
2.	18	15	09	37	48	14	04	16	29	16	10	03	14	62
3.	21	16	06	61	41	08	—07	13	26	19	09	04	05	76
4.	00	11	13	24	19	03	04	23	16	08	07	—07	23	27
6.	19	32	02	06	08	04	05	10	03	56	02	00	—10	49
7.	21	33	00	—04	09	00	03	01	05	39	15	07	04	35
8.	36	58	—02	—10	14	09	02	12	03	31	24	02	05	67
9.	—01	24	35	20	08	59	13	26	08	05	01	05	—04	67
10.	20	25	28	34	20	09	18	23	11	09	26	—10	02	52
11.	14	14	59	44	22	14	20	08	02	14	20	—09	13	77
12.	01	21	20	22	18	08	25	53	13	04	17	00	—06	56
13.	05	12	02	30	06	11	11	29	05	25	28	—09	—02	37
14.	03	24	07	20	16	—06	09	50	28	12	30	—10	04	58
15.	05	16	44	45	25	11	16	29	04	00	14	03	07	64
17.	08	28	14	08	01	23	06	18	36	06	00	—01	21	37
18.	00	11	09	09	60	14	19	03	03	16	04	13	—04	49
19.	07	19	14	63	15	16	07	27	16	16	—04	13	—04	65
20.	05	26	—02	05	13	—05	18	32	23	20	30	23	03	47
21.	19	02	15	41	34	09	21	08	17	11	11	14	00	47
23.	32	06	17	14	—10	05	29	13	—04	24	20	23	16	44
24.	03	06	61	23	22	05	09	22	07	22	06	14	06	62
25.	18	11	12	28	03	13	64	12	12	19	07	11	05	65
26.	22	16	07	31	04	05	55	16	19	21	01	17	—02	61
27.	10	02	20	03	08	58	02	05	25	07	30	07	20	61
28.	33	29	14	10	31	16	12	16	23	—03	26	—05	—07	51
29.	07	19	07	25	16	—02	01	26	11	03	37	11	21	41
30.	24	18	07	—06	45	16	10	27	01	00	20	25	20	54
31.	04	25	34	34	11	20	12	07	19	—10	14	18	16	49
32.	00	46	—05	23	12	16	12	—10	06	28	06	—02	20	46
33.	15	38	11	25	16	—03	06	26	12	32	24	11	32	63
34.	27	18	01	—03	05	13	07	05	00	16	50	18	—02	44
35.	26	19	11	—04	11	—04	09	13	03	11	46	16	03	41
36.	09	05	04	02	17	48	02	07	16	26	05	21	—12	44
37.	29	14	16	—03	42	—03	14	23	23	06	16	05	08	46
38.	32	09	08	01	24	09	24	—09	08	17	09	66	13	74
39.	23	11	10	12	22	—09	14	16	19	09	08	14	—11	27
40.	11	22	—05	30	23	17	21	22	—08	24	19	03	09	44
41.	—01	04	24	50	00	06	—09	03	05	04	11	32	14	47
42.	34	09	04	06	38	00	30	34	10	15	11	05	21	57
51.	16	08	40	30	11	14	04	20	01	19	04	24	—06	46
52.	—06	—06	26	58	20	12	06	16	08	06	06	08	03	52
53.	11	53	16	14	00	19	03	05	06	16	04	14	—04	43
54.	77	04	02	—04	08	00	12	02	06	21	05	00	03	67
55.	80 ✓	08	01	07	16	—03	04	06	12	09	02	08	—03	70
59.	45	—04	07	30	48	04	00	12	15	05	06	01	07	58
63.	06	14	27	23	02	14	15	18	01	30	21	14	02	38

*Decimal points omitted.

45 57 51 83 68 37 51 58 37 52 49 36 26

TABLE 4
Rotated (Orthogonal) Factor Loadings (Solution II)*

Test No.	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1.	03	00	14	13	04	26	10	18	43	27	00	26
2.	20	15	16	41	36	06	00	13	37	25	05	14
3.	14	15	18	60	26	—01	08	04	34	23	07	04
4.	01	07	14	26	17	—04	01	28	24	12	—05	05
6.	06	41	22	07	00	08	12	06	—03	42	05	15
7.	08	32	09	—03	01	00	07	12	05	36	21	16
8.	17	57	00	—05	07	05	06	25	11	26	29	23
9.	—07	26	26	20	06	59	15	09	29	—11	03	11
10.	18	25	27	30	21	11	17	29	27	04	17	—04
11.	12	14	58	26	28	12	12	18	37	03	17	—11
12.	—01	15	21	32	19	16	24	43	18	—05	02	17
13.	05	17	15	34	11	17	17	29	00	19	14	—03
14.	02	16	09	33	13	06	11	57	19	13	05	08
15.	07	11	44	40	25	10	11	23	33	—09	10	05
17.	02	19	01	11	01	23	04	24	43	12	—05	12
18.	01	12	14	15	51	00	08	—08	20	17	11	24
19.	02	19	23	60	15	12	14	10	29	12	02	19
20.	—04	09	03	18	05	04	20	38	16	19	23	32
21.	22	01	21	37	27	04	18	01	28	14	12	09
23.	24	03	26	07	—04	11	32	16	06	15	24	17
24.	01	01	62	17	19	13	01	18	29	06	06	16
25.	13	05	22	17	13	13	62	09	25	14	07	12
26.	16	08	20	23	03	06	57	07	25	17	03	18
27.	09	01	02	05	09	59	00	16	32	13	24	02
28.	29	26	03	15	25	12	09	20	28	08	25	10
29.	06	06	07	33	13	—01	02	38	19	08	27	11
30.	24	19	09	06	59	05	00	16	15	07	20	09
31.	—04	07	21	28	10	14	09	12	50	—10	21	10
32.	—09	42	03	17	10	—01	15	01	26	31	16	04
33.	04	29	20	28	08	—05	08	38	23	30	21	22
34.	19	12	01	02	08	18	11	21	01	16	50	13
35.	18	09	09	01	10	04	09	32	07	12	44	15
36.	04	13	—03	14	15	47	00	—11	10	24	12	24
37.	27	09	09	05	34	—04	05	28	20	15	08	26
38.	18	—09	13	03	09	02	19	—16	25	19	38	57
39.	19	06	13	15	09	—04	13	12	18	10	09	19
40.	08	26	12	34	26	09	23	15	03	19	17	14
41.	—04	—08	30	45	—08	05	—05	—01	26	01	24	08
42.	32	11	10	14	36	—04	22	31	14	18	00	28
51.	11	08	47	27	04	20	05	03	17	01	12	20
52.	00	—06	36	55	16	11	06	02	23	01	04	—02
53.	—02	50	21	09	04	14	09	00	26	10	21	—03
54.	70	20	02	—09	04	02	14	04	03	24	08	11
55.	73	20	02	05	03	—03	05	02	14	14	09	17
59.	51	07	10	33	35	—04	—05	03	21	14	02	10
63.	00	16	40	20	06	10	18	19	07	11	20	11

*Decimal points omitted.

PSYCHOMETRIKA—VOL. 17, NO. 1
MARCH, 1952

THE EFFECT OF DIFFICULTY AND CHANCE SUCCESS ON ITEM-TEST CORRELATION AND ON TEST RELIABILITY*

LYNNETTE B. PLUMLEE
EDUCATIONAL TESTING SERVICE

An equation is derived for predicting the effect of chance success, relative to item difficulty, on item-test correlation. The values predicted by this equation and by equations derived by Guilford and Carroll for predicting the effect of chance success on item difficulty and test reliability are compared with empirical values in an experiment which used identical test items in multiple-choice and answer-only form.

Introduction

The "multiple-choice" type of test, in which answer options are supplied, frequently has been objected to on the grounds that an examinee who does not know the answer to any of the items in a test can make a substantial score by pure guesswork and that his score therefore does not represent his true knowledge. In support of the multiple-choice test, on the other hand, it has been argued that this type of test can be made both a more effective and a more efficient measuring instrument than the "answer-only" type of test, in which no answer options are given and which thus require the examinee to supply his own answers. Since wrong answer options can be restricted to answers resulting from certain types of errors, the multiple-choice item can be directed towards testing specific types of errors. The multiple-choice test places the burden of decision regarding the correctness of an answer on the examinee rather than on the scorer, and greater consistency in scoring is thus possible in this type of test than in the "answer-only" test.

One aspect of the problem is investigated in this paper: What is

*Condensation of a dissertation presented in partial fulfillment of the requirements for the Ph.D. degree to the University of Chicago. Grateful acknowledgment is made to Professor Harold Gulliksen for his guidance as thesis advisor and to Professor L. L. Thurstone and Dr. D. W. Fiske of the University of Chicago who served as members of the thesis committee. The author is also indebted to Professor S. S. Wilks for review of the derivations and development of statistical tests used in the thesis, to Dr. L. R Tucker for technical advice, and to Dr. W. G. Mollenkopf for critical comments on the derivations and interpretations. The writer expresses appreciation to the Educational Testing Service for making available its technical facilities, and to the University of Chicago for the flexible administrative arrangement which made this thesis possible.

the theoretical effect of chance success on item-test correlation and on test reliability, and to what extent is this theoretical expectation borne out in practice?

Several previous investigations (1-13, 15) have been concerned with the prediction of the effect of chance success on item difficulty, inter-item correlation, item-test correlation, and test reliability.

In 1936, Guilford (3) presented a formula for the proportion of examinees who know the answer to an item, ${}_cp$, as a function of the proportion of examinees who answer the item correctly in multiple-choice form, p, and the number of answer options, n:

$$ {}_cp = \frac{np}{n-1} - \frac{1}{n-1}. \tag{1} $$

This formula assumes that all answer options are equally attractive to the examinee who does not know the correct answer.

Carroll (1) considered the combined effect of item difficulty and chance success on the Pearsonian correlation between items or between sets of items. Using the binomial distribution of chance failure (number-not-right) scores, he arrived at the following formulas for mean, standard deviation, inter-test correlation, and inter-item correlation, respectively:

$$ \bar{E}' = W\bar{E}\,, \tag{2} $$

$$ \sigma_{E'} = \sqrt{W^2\sigma_E{}^2 + RW\bar{E}}\,, \tag{3} $$

$$ r_{E_1'E_2'} = \frac{W r_{E_1E_2}\,\sigma_{E_1}\sigma_{E_2}}{\sqrt{W^2\,\sigma_{E_1}^2\,\sigma_{E_2}^2 + RW\bar{E}_1\,\sigma_{E_2}^2 + RW\bar{E}_2\,\sigma_{E_1}^2 + R^2\bar{E}_1\bar{E}_2}}\text{, and} \tag{4} $$

$$ r_{1'2'} = \frac{Wq_1(1-q_2)}{\sqrt{q_1q_2(1-Wq_1)(1-Wq_2)}}\,, \tag{5} $$

where E represents the failure score, W is the probability of failure and R is the probability of success in multiple-choice form, a prime denotes a multiple-choice test, absence of a prime denotes an answer-only test, subscripts 1 and 2 denote two different items or tests, and q is the probability of failing a given item in answer-only form. (Notations used by the present author have been substituted for some of those used by Carroll.)

It is the plan of the present paper to develop the equation for predicting the effect of chance success on item-test correlation (biserial coefficient of correlation) and to compare the values predicted by this equation and equations (1) and (4) with values actually obtained in an experiment which used identical test items in multiple-choice and answer-only form.

The Prediction Equation

In deriving the equation, the following assumptions will be made:

1. that every examinee attempts every item in the test,
2. that every examinee who knows the correct answer to an item answers the item correctly in both multiple-choice and answer-only form,
3. that every examinee who does not know the correct answer to an item answers the item incorrectly in answer-only form and chooses from among the options on a basis of chance alone in multiple-choice form, and
4. that the number of options per item is the same for all items in the multiple-choice form of the test.

Although the first three assumptions are rarely if ever borne out in an actual test situation, they are necessary in deriving relationships between a hypothetical test on which results are not influenced by chance success and one in which chance success is fully operative.

The following notations will be used:

Multiple-Choice Form of Test	Answer-Only Form of Test	
R		Probability of answering an item correctly on the basis of chance alone.
W		Probability of answering an item incorrectly on the basis of chance alone, where $W = 1 - R$.
h	h	Number of items in the test.
x'	x	Score (number of items correct) on the test.
x'_k	x_k	Score of kth individual.

Multiple-Choice Form of Test	Answer-Only Form of Test	
a'_{kj}	a_{kj}	Score of *k*th individual on *j*th item, where $a = 0$ or 1 depending on whether the item is answered incorrectly or correctly.
t	t	Number of individuals who take the test.
$t_{+'}$	t_+	Number of individuals who answer a given item correctly.
	t_-	Number of individuals who answer a given item incorrectly.
$p' = \frac{t_{+'}}{t}$	$p = \frac{t_+}{t}$	Proportion of all examinees who answer a given item correctly.
$M_{x'}$	M_x	Mean score on the test.
$M_{+'}$	M_+	Mean score of those individuals who answer a given item correctly.
	M_-	Mean score of those individuals who answer a given item incorrectly.
$\sigma_{x'}$	σ_x	Standard deviation of scores on the test.
z'	z	Ordinate of the normal probability curve at a point to the left of which lies a proportion p' or p of the area under the curve.
r_{bis}'	r_{bis}	Biserial correlation between a given item and total test, where

$$r_{\text{bis}} = \frac{M_+ - M_x}{\sigma_x} \cdot \frac{p}{z} \tag{6}$$

and

$$r_{\text{bis}}' = \frac{M_{+'} - M_{x'}}{\sigma_{x'}} \cdot \frac{p'}{z'}. \tag{7}$$

An expression can be found for r_{bis}' in terms of r_{bis} if we know the values of $M_{+'}$, $M_{x'}$, $\sigma_{x'}$, and p' in terms of the corresponding answer-only statistics.

Guilford expressed p in terms of p' as shown in equation (1). An equivalent formula for p' in terms of p is

$$p' = Wp + R\,. \tag{8}$$

Carroll's formulas for the mean, (2), and standard deviation, (3), can be expressed in terms of number-right scores as follows:

$$M_{x'} = WM_x + Rh, \text{ and} \tag{9}$$

$$\sigma_{x'} = \sqrt{W^2\,\sigma_x^{\,2} + RW(h - M_x)}. \tag{10}$$

It will be noted that as $\sigma_x^{\,2}$ becomes very large relative to $h - M_x$ or as W approaches 1

$$\lim \frac{\sigma_{x'}}{\sigma_x} = W. \tag{11}$$

If, for an answer-only form of a standardized test, the mean is equal to about one-half the range and the standard deviation is equal to about one-fifth the range, where the range approximates the total number of items, then the standard deviation of the multiple-choice form of the test, $\sigma_{x'}$, will theoretically approach $W\sigma_x$ as the number of items becomes large or as the number of answer options becomes large.

To find $M_{+'}$ in (7), we note that the sum of answer-only scores of those t_+ individuals who answer item j correctly in answer-only form is

$$\sum_{k=1}^{t} x_k a_{kj}\,. \tag{12}$$

From (9) it will be seen that

$$\sum_{k=1}^{t} x'_k = W \sum_{k=1}^{t} x_k + Rht. \tag{13}$$

Therefore, the sum of multiple-choice scores of the t_+ individuals is

$$W \sum_{k=1}^{t} x_k a_{kj} + Rht_+. \tag{14}$$

The sum of answer-only scores of those t_- individuals who answer item j incorrectly in answer-only form is

$$\sum_{k=1}^{t} x_k(1-a_{kj})\,. \tag{15}$$

Of these t_- individuals, a fraction R, selected at random, will answer item j correctly in multiple-choice form, and the sum of multiple-choice scores for the latter group will be

$$RW\sum_{k=1}^{t} x_k(1-a_{kj}) + R^2ht_-. \tag{16}$$

Therefore, the sum of multiple-choice scores for all persons who answer item j correctly in multiple-choice form will be the sum of (14) and (16):

$$\begin{aligned}\sum_{k=1}^{t} x'_k a'_{kj} = W\sum_{k=1}^{t} x_k a_{kj} + RW\sum_{k=1}^{t} x_k \\ - RW\sum_{k=1}^{t} x_k a_{kj} + Rht_+ + R^2ht_-.\end{aligned} \tag{17}$$

Since

$$t_- = t - t_+ \quad \text{and} \quad W = 1 - R\,,$$

$$\sum_{k=1}^{t} x'_k a'_{kj} = W^2\sum_{k=1}^{t} x_k a_{kj} + RW\sum_{k=1}^{t} x_k + RWht_+ + R^2ht. \tag{18}$$

Since

$$\sum_{k=1}^{t} x_k a_{kj} = t_+M_+ \quad \text{and} \quad \sum_{k=1}^{t} x_k = tM_x\,,$$

then

$$\sum_{k=1}^{t} x'_k a'_{kj} = W^2t_+M_+ + RWtM_x + RWht_+ + R^2ht. \tag{19}$$

Dividing both sides of (19) by $t_{+'}$ and factoring out t in the right-hand side of the equation, we obtain

$$M_{+'} = \frac{t}{t_{+'}}\left(W^2pM_+ + RWM_x + RWhp + R^2h\right). \tag{20}$$

Noting from (9) and (8) that

$$\begin{aligned}M_{x'} = \frac{p'}{p'}\left(WM_x + Rh\right) = \frac{1}{p'}\left(Wp + R\right)\left(WM_x + Rh\right) \\ = \frac{1}{p'}\left(W^2pM_x + RWM_x + RWhp + R^2h\right)\end{aligned} \tag{21}$$

and that

$$\frac{t}{t_{+'}} = \frac{1}{p'},$$

we have

$$M_{+'} - M_{x'} = \frac{1}{p'}\,(W^2pM_+ - W^2pM_x) = \frac{1}{p'}\,[W^2p(M_+ - M_x)]. \quad (22)$$

Since from (6)

$$M_+ - M_x = \frac{r_{\text{bis}}\sigma_x z}{p}, \quad (23)$$

$$r_{\text{bis}}' = W^2 \cdot \frac{\sigma_x}{\sigma_{x'}} \cdot \frac{z}{z'} \cdot r_{\text{bis}}. \quad (24)$$

Since the relationship z/z' is constant for any given values of p and R, this ratio may be easily obtained from a table or graph prepared for such a purpose. F. M. Lord has shown that, as p increases from 0 to 1, z/z' increases with decreasing acceleration from 0 to a limit, $1/W$.

It will be noted that for the conditions under which $\sigma_{x'}$ approaches $W\sigma_x$

$$\lim r_{\text{bis}}' = W \cdot \frac{z}{z'} \cdot r_{\text{bis}}. \quad (25)$$

Also, as p approaches 1, under the conditions for (25), r_{bis}' approaches r_{bis}.

Carroll's formula for the reliability of a test in multiple-choice form, (4), may be expressed in terms of number-right scores as follows:

$$r_{x_1'x_2'} = \frac{W r_{x_1x_2}\,\sigma_{x_1}\sigma_{x_2}}{\sqrt{W\sigma_{x_1}^2 + R(h - M_{x_1})}\,\sqrt{W\sigma_{x_2}^2 + R(h - M_{x_2})}}. \quad (26)$$

It will be noted that the reliability of the multiple-choice form of a given test will theoretically always be less than the reliability of the answer-only form of the same test. However, under the conditions for which $\sigma_{x'}$ approaches $W\sigma_x$, the multiple-choice reliability will theoretically approach the answer-only reliability as a limit.

Comparison of Observed Data and Predicted Values

In order to determine the extent to which the predicted effects of chance success are obtained in practice, a series of mathematics tests was designed, employing the same items in answer-only and multiple-choice form.

The Tests

Four sections of 36 items each were prepared; all sections were planned to be parallel in difficulty, discriminative power, solution time, and subject matter. Estimates of item difficulty and discriminative power were obtained from statistics from previous uses of the items in answer-only form. Estimates of subject-matter equivalence and time required for solution were subjective. Equivalence of subject matter, which included algebra, geometry, and trigonometry, was considered rather broadly.

Each of the four sections was prepared both in multiple-choice form (with five answer options) and in answer-only form. Two alternative methods were considered in selecting answer options. Either the true chance situation might be approached as nearly as possible by selecting answer options within close range of or very similar to the correct answer, or the practical testing situation might be approached and answer options be selected on the basis of their appeal to the examinee who does not know the correct answer. Since the latter approach was felt to be more meaningful for the purposes of test construction, options were selected by the usual technique, which includes using answers reached by popular wrong methods of solution. In some instances, tallies of the frequencies of actual answers given by examinees to an item in answer-only form were used in selecting answer options for the multiple-choice form of the item.

The four 36-item sections in multiple-choice form may be designated as M1, M2, M3, and M4, and the same sections in answer-only form as A1, A2, A3, and A4, respectively. For purposes of analysis, the items in sections M1, M2, A1, and A2 will be referred to as Set I items; those in sections M3, M4, A3, and A4, as Set II items.

An additional set of 16 mathematics items, alternating multiple-choice and answer-only, was administered to all examinees as a basis for checking the equivalence of the population groups.

Four tests were then arranged as follows:

	Test W	Test X	Test Y	Test Z
Part I	Set of 16 items, common to all tests, with even-numbered items in multiple-choice form and odd-numbered items in answer-only form.			
Part II	Item Set I Section M1	Item Set II Section A3	Item Set II Section M3	Item Set I Section A1
Part III	Item Set I Section M2	Item Set II Section A4	Item Set II Section M4	Item Set I Section A2
Part IV	Item Set II Section A3	Item Set I Section M1	Item Set I Section A1	Item Set II Section M3
Part V	Item Set II Section A4	Item Set I Section M2	Item Set I Section A2	Item Set II Section M4

In order to determine how many examinees actually looked at all items, an easy item, which it was expected would be tried by virtually all of those examinees who reached it, was placed at the end of each section. Fifteen minutes working time was allowed for Part I and thirty-five minutes for each of the other parts.

The Sample of Individuals Tested

The tests were administered to a sample of approximately 560 male examinees of college freshman level or higher. The four test booklets were distributed in successive order. The analysis was based on 138 cases for Test W and on 139 cases for each of the other tests.

Scoring

When each test was scored, "total-number-correct" scores were obtained for each part separately, for Parts II and III combined, and for Parts IV and V combined. Thus, for each test and each examinee there were a score on the part common to all four tests, a score on the multiple-choice items, and a score on the answer-only items.* The easy last item in each section was not scored or included in the further analysis.

Item Analysis

For each item in Set I and Set II the proportion of examinees

*Item 17 in Set I proved to have two defensible answers among the answer options in multiple-choice form. In scoring and in subsequent analyses, both answers were considered correct, but the item was treated as a five-choice item for prediction purposes. The obtained proportion correct in multiple-choice form was .88 for Part II and .83 for Part IV; in answer-only form, .63 and .76. The two obtained biserial correlations in multiple-choice form were .55 and .38; in answer-only form, .51 and .63. The item's average multiple-choice difficulty approached that predicted for a two-choice item, but its average biserial was more nearly that predicted for a five-choice item.

that marked the correct answer choice was computed. This proportion had as its base the number of examinees who answered the given item or a subsequent item in the same part. Since the number of candidates who completed the scored items in the different parts varied from 32 to 109, the aim of a true power test was not met, and many of the proportion-correct figures are not based on the total number of examinees. For each answer-only item, the predicted multiple-choice proportion correct was computed from the answer-only proportion correct, using equation (8).

Also, for each item in Sets I and II, the biserial coefficient of correlation was computed, with the total score on the item set of which the item was a part as the criterion. Hence, multiple-choice items were analyzed against the multiple-choice score as the criterion, and answer-only items were analyzed against the answer-only score as the criterion. For each answer-only item the predicted multiple-choice biserial coefficient of correlation was computed from the answer-only biserial coefficient, using equation (24).

Analysis of Data

In order to determine whether the item analysis and reliability data for multiple-choice differed from that predicted for multiple-choice from answer-only more than could be accounted for by chance, certain comparisons of observed values with the corresponding predicted values were made. In the presentation of these comparisons,

$\hat{p}$ = the observed sample value of the answer-only proportion correct,

$\hat{p}'$ = the observed sample value of the multiple-choice proportion correct,

$\hat{r}$ = the observed sample value of the answer-only biserial,

$\hat{r}'$ = the observed sample value of the multiple-choice biserial,

$\tilde{p}'$ = the value of the multiple-choice proportion correct predicted by equation (8) from the observed answer-only proportion correct, and

$\tilde{r}'$ = the value of the multiple-choice biserial predicted by equation (24) from the observed answer-only biserial.

The subscript 1 indicates that the item appeared in Part II or III, the subscript 2, that it appeared in Part IV or V.

Group equivalence. To determine whether the item statistics and reliability values can be compared directly from test to test, the four groups were compared on the basis of their scores on Part I.

An analysis of variance of the differences among the means of the various groups indicated that they were not significantly different. ($F = 1.3$, d.f. $= 3$ and 551, $p > 5\%$.)

Item difficulty. The regression line of $\hat{p}'$ on $\hat{p}$ was compared with the theoretical relationship between multiple-choice and answer-only proportion correct for five-choice items:

$$p' = .8p + .2 \,. \tag{27}$$

To test whether the regression parameters of the observed difficulty values differed from the theoretical parameters, given by equation (27), more than would be expected by chance, the 95% confidence limits of the true regression parameters were computed for each regression line. (See Table 1.) The significance tests used are those described by Wilks in (16) for the regression coefficient and in (17) for the regression intercept. The hypothesis that the theoretical parameters fall within the 95% confidence limits of the true parameter values is supported for seven of the eight regression lines. In the case of the regression of $\hat{p}'_2$ on $\hat{p}_2$ for Item Set I, the parameters are outside the 95% confidence range in the direction of the parameters of the relationship $p' = p$. However, since the tendency of the obtained Set I regression coefficients to exceed the theoretical coefficients is offset by the tendency of the obtained Set II regression coefficients to fall below the theoretical, there seems to be no basis for attaching significance to the direction of the difference. No variation in the test content or administration was found which would account for the noted difference between Item Sets, although the probability of this difference occurring by chance (according to a simple signs test) would be only .008.

As an additional check on the relative difficulty of items in the two answer forms, the means and standard deviations of the proportion correct, observed and predicted by equation (8), are shown in Table 2. It will be noted that $\tilde{p}'_1$ means for both sets correspond closely to $\hat{p}'_1$ and $\hat{p}'_2$ means, but that $\tilde{p}'_2$ values are consistently higher.

TABLE 1*
Regression of p' on p

	Theoretical	Obtained Values			
	Values	$\hat{p}'_1$ on $\hat{p}_1$	$\hat{p}'_2$ *on* $\hat{p}_2$	$\hat{p}'_2$ on $\hat{p}_1$	$\hat{p}'_1$ on $\hat{p}_2$
Item Set I					
Regression coefficient and 95% confidence limits	.80	.807 (±.094)	.866 (±.064)	.836 (±.089)	.820 (±.082)
Regression intercept and 95% confidence limits	.20	.201 (±.051)	.124 (±.038)	.181 (±.048)	.154 (±.048)
Correlation coefficient		.895	.953	.910	.920
Item Set II					
Regression coefficient and 95% confidence limits	.80	.769 (±.099)	.765 (±.077)	.780 (±.075)	.738 (±.106)
Regression intercept and 95% confidence limits	.20	.224 (±.055)	.193 (±.046)	.222 (±.042)	.204 (±.064)
Correlation coefficient		.876	.919	.924	.852

*$\hat{p}_1$ and $\hat{p}'_1$ = observed answer-only and multiple-choice proportions correct, respectively, for Parts II and III.

$\hat{p}_2$ and $\hat{p}'_2$ = observed answer-only and multiple-choice proportions correct, respectively, for Parts IV and V.

TABLE 2*
Comparison of Item Difficulty Means and Standard Deviations

Statistic	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}'_1$	$\hat{p}'_2$	$\tilde{p}'_1$	$\tilde{p}'_2$
Item Set I						
Mean	.473	.523	.583	.577	.580	.619
σ	.256	.259	.231	.235	.207	.208
Item Set II						
Mean	.494	.542	.603	.607	.597	.632
σ	.259	.263	.228	.219	.205	.205

*$\hat{p}_1$ and $\hat{p}'_1$ = observed answer-only and multiple-choice proportions correct, respectively, for Parts II and III.

$\hat{p}_2$ and $\hat{p}'_2$ = observed answer-only and multiple-choice proportions correct, respectively, for Parts IV and V.

$\tilde{p}'_1$ = multiple-choice proportion correct predicted by equation (8) from observed answer-only proportion correct for Parts II and III.

$\tilde{p}'_2$ = multiple-choice proportion correct predicted by equation (8) from observed answer-only proportion correct for Parts IV and V.

This difference, which reflects that between $\hat{p}_2$ and $\hat{p}_1$, may indicate the influence of a practice effect. However, the average observed mul-

TABLE 3*
Comparison of Biserial Correlation Values for Items in Multiple-Choice and Answer-Only Form

	Observed Values, Same Answer Form		Observed Values, Different Answer Form				Observed Values with Predicted Values			
	$\hat{r}'_2$ and $\hat{r}'_1$	$\hat{r}_2$ and $\hat{r}_1$	$\hat{r}'_1$ and $\hat{r}_1$	$\hat{r}'_2$ and $\hat{r}_2$	$\hat{r}'_2$ and $\hat{r}_1$	$\hat{r}'_1$ and $\hat{r}_2$	$\hat{r}'_1$ and $\tilde{r}'_1$	$\hat{r}'_2$ and $\tilde{r}'_2$	$\hat{r}'_2$ and $\tilde{r}'_1$	$\hat{r}'_1$ and $\tilde{r}'_2$
Item Set I										
σ_{est} of first value from second	.128	.152	.147	.149	.135	.151	.143	.131	.124	.137
σ_{est} of second value from first	.134	.152								
σ_e of estimate from line of 45° slope through origin	.150	.188	.189	.212	.172	.205	.181	.151	.160	.153
Difference between means of values (first minus second)	.005	.019	-.058	-.072	-.053	-.077	.076	.055	.081	.050
Correlation between values	.53	.31	.36	.18	.45	.29	.42	.50	.57	.50
Item Set II										
σ_{est} of first value from second	.136	.145	.172	.148	.147	.172	.163	.138	.130	.168
σ_{est} of second value from first	.159	.151								
σ of estimate from line of 45° slope through origin	.188	.178	.254	.234	.217	.275	.180	.170	.166	.202
Difference between means of values (first minus second)	.058	.040	-.110	-.093	-.053	-.151	.027	.031	.085	-.026
Correlation between values	.38	.46	.10	.05	.12	.06	.34	.36	.48	.22

*$\hat{r}_1$ and $\hat{r}'_1$ = observed answer-only and multiple-choice biserials, respectively, for Parts II and III.

$\hat{r}_2$ and $\hat{r}'_2$ = observed answer-only and multiple-choice biserials, respectively, for parts IV and V.

$\tilde{r}'_1$ = multiple-choice biserial predicted by equation (24) from observed answer-only biserial for Part II or Part III.

$\tilde{r}'_2$ = multiple-choice biserial predicted by equation (24) from observed answer-only biserial for Part IV or Part V.

tiple-choice proportion correct does not seem to be increased by practice. The variance of the observed multiple-choice proportion correct values lies consistently between the variance of the predicted multiple-choice proportion correct and the observed answer-only proportion correct.

Item-test correlation. The correlations between $\hat{r}'$ and $\tilde{r}'$ were sufficiently low to rule out a satisfactory comparison of the obtained regression coefficients and the coefficients of the theoretical relationship by the method used in analyzing item difficulty. In order to determine whether these low correlations between the observed and predicted biserials represented a significantly low relationship or whether they were as high as could be expected from the extent of agreement between two sets of observed biserials for the same items in the same answer form, correlations were obtained between $\hat{r}_2$ and $\hat{r}_1$ and between $\hat{r}'_2$ and $\hat{r}'_1$.

In order to determine further whether the values obtained by the application of equation (24) to the observed answer-only values were in closer agreement with the observed multiple-choice biserials than were the observed answer-only values, correlations were also obtained between $\hat{r}'$ and $\hat{r}$.

Since there were too few extreme values of the biserial coefficient to stabilize regression lines, a direct comparison of correlation

TABLE 4*
Reliability of Tests

Statistic	Test W	Test X	Test Y	Test Z
Number of cases	138	139	139	139
Mean score				
Observed answer-only	37.7	31.6	36.2	31.1
Observed multiple-choice	37.8	39.8	39.7	42.2
Standard deviation of scores				
Observed answer-only	11.3	10.6	10.9	10.6
Observed multiple-choice	10.2	10.4	9.6	10.3
Reliability				
Observed answer-only*	.84	.80	.85	.81
Observed Multiple-choice*	.81	.76	.72	.81
Predicted multiple-choice†	.75	.69	.75	.69

*Correlation between two parallel separately timed parts.
†Correlation predicted by equation (26) from the observed answer-only correlation between parts.

coefficients was not considered meaningful, and hence the following standard errors of estimate were computed:

1. the usual standard error of estimate from the best-fit regression line, and
2. the standard error of estimate from a line through the origin with a 45° slope, $\sqrt{\sigma_X^2 + \sigma_Y^2 - 2r_{XY}\sigma_X\sigma_Y + (\bar{X} - \bar{Y})^2}$, in order to test the hypothesis that the compared values are truly equal. Table 3 shows these errors of estimate, the correlations, and the differences between mean values.

The values in Table 3 seem to indicate that the lack of agreement between observed and predicted values is no greater than that between two sets of observed values for items in the same answer form. It will be noted, however, that the standard error of estimate where observed and predicted multiple-choice biserials are being compared is consistently lower than the corresponding standard error of estimate where observed multiple-choice values are being compared with observed answer-only values. This would indicate that the values obtained by the application of equation (24) to the observed answer-only biserials are more in agreement with the observed multiple-choice values than are the observed answer-only values.

From the differences between mean values, it is seen that the observed answer-only biserials are greater than the observed multiple-choice biserials for the same items by an average of .08 point, but that the predicted multiple-choice biserials are an average of .05 point lower than the observed multiple-choice biserials. It would appear, therefore, that on the average equation (24) over-corrects.

Test-reliability. The reliability values reported in Table 4 for observed answer-only and multiple-choice scores are the correlations between two separately timed parts, Part II and Part III or Part IV and Part V. The predicted multiple-choice reliabilities were computed from the observed answer-only reliabilities using equation (26). It will be noted, then, that each of the reported reliability values is for a test of 36 items.

From Table 4 it is seen that the observed multiple-choice reliability values tend to fall between the observed answer-only and predicted multiple-choice values. In order to test whether the differences among the observed reliability figures were significant, each reliability coefficient was transformed to a Fisher z-value. A chi-square test (14) was then applied to the eight values for observed multiple-choice and answer-only. Since the obtained chi-square for the eight values was 13.6 and the probability under the null hypothesis is .05 that, for seven degrees of freedom, chi-square will exceed 14.1, it was concluded that the differences among the observed multiple-choice and answer-only reliability values are not significant.

To test whether or not the differences between observed and predicted multiple-choice reliabilities were significant, the 95% confidence limits of the true multiple-choice Fisher z-values were computed. When predicted and observed multiple-choice reliabilities are compared for the same group but different item sets, the hypothesis that the predicted multiple-choice z-value falls within the 95% confidence limits of the true Fisher z-value is supported in three out of four cases. When the set of items is held constant but the groups differ, the predicted multiple-choice z-value is within the 95% confidence limits of the true multiple-choice Fisher z-value in six out of eight cases.

In the absence of any clear-cut evidence to the contrary, it may be concluded that the observed multiple-choice reliability appears to lie between the observed answer-only value and the multiple-choice value which is predicted from the observed answer-only value by equation (26). The observed multiple-choice value does not differ consistently from either the observed answer-only value or the predicted multiple-choice value.

Discussion

Although, in most instances, the differences between observed and predicted multiple-choice statistics do not seem to be significant, there appears to be a fairly consistent tendency for the differences, slight though they may be, to be in the direction of the answer-only statistics, away from the predicted multiple-choice statistics.

Some explanation for this discrepancy between the theoretical and observed statistics may be found by examining the extent to which the first three assumptions were met.

As indicated in the description of the item analysis, although the aim was to set a test such that nearly every examinee would attempt every item, this aim was not met. To test whether "drop-out" as such influenced the results, the last twelve items in each part were plotted in a distinctive color on the correlation plots of item difficulty and biserial correlation. However, there seemed to be no definite pattern distinguishing these items from the others.

Since examinees may make careless errors or may omit items which they are able to solve, the assumption that an examinee who knows the correct answer to an item answers the item correctly in both multiple-choice and answer-only form would not be expected to hold in the current experimental situation. There seems to be no objective way of determining exactly to what extent this assumption was met.

Although the assumption that an examinee who does not know the correct answer to an item answers the item incorrectly in answer-only form probably held, no attempt was made to meet the assumption that such an examinee will answer the item according to chance in multiple-choice form, since it was felt desirable to match the current test construction practice of including among the options for a multiple-choice item those incorrect answers which will be obtained by a large number of examinees on the basis of a wrong solution. Logically, the result of such a procedure is to eliminate the operation of chance for those examinees who use a popular wrong method of solution. Hence, the failure to utilize wrong answer options which would be equally attractive to an examinee who does not know the correct answer might alone account for the tendency of observed multiple-choice values to differ from predicted multiple-choice values in the direction of answer-only values.

One factor which may have operated to keep the proportion answering correctly high in multiple-choice form in spite of a reduced chance element is the check which examinees have on their answers.

Although these explanations cannot be assumed from the data presented in the current study, the evidence does seem to indicate that item-test correlation and test reliability may not be as adversely affected by the multiple-choice form as has frequently been assumed.

REFERENCES

1. Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika,* 1945, **10**, 1-19.
2. Denney, H. R., and Remmers, H. H. Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula, II. *J. educ. Psychol.*, 1940, **31**, 699-704.
3. Guilford, J. P. The determination of item difficulty when chance success is a factor. *Psychometrika,* 1936, **1**, 259-264.
4. Horst, Paul. The chance element in the multiple choice test item. *J. gen. Psychol.*, 1932, **6**, 209-211.
5. Horst, Paul. The difficulty of a multiple-choice test item. *J. educ. Psychol.*, 1933, **24**, 229-232.
6. Horst, Paul. The difficulty of multiple choice test item alternatives. *J. exp. Psychol.*, 1932, **15**, 469-472.
7. Johnson, A. P. An index of item validity providing a correction for chance success. *Psychometrika,* 1947, **12**, 51-58.

8. Lord, F. M. Reliability of multiple-choice tests as a function of number of choices per item. *J. educ. Psychol.*, 1944, **35**, 175-180.

9. Remmers, H. H., and Adkins, R. M. Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula, VI. *J. educ. Psychol.*, 1942, **33**, 385-390.

10. Remmers, H. H., and Ewart, Edwin. Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula, III. *J. educ. Psychol.*, 1941, **32**, 61-66.

11. Remmers, H. H., and House, J. M. Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula, IV. *J. educ. Psychol.*, 1941, **32**, 372-376.

12. Remmers, H. H., Karslake, Ruth, and Gage, N. L. Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula, I. *J. educ. Psychol.*, 1940, **31**, 583-590.

13. Remmers, H. H., and Sageser, H. W. Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown formula, V. *J. educ. Psychol.*, 1941, **32**, 445-451.

14. Rider, P. R. An introduction to modern statistical methods. New York: John Wiley & Sons, Inc., 1939.

15. Votaw, D. F. Notes on validation of test items by comparison of widely spaced groups. *J. educ. Psychol.*, 1934, **25**, 185-191.

16. Wilks, S. S. Elementary statistical analysis. Princeton, New Jersey: Princeton Univ. Press, 1948.

17. Wilks, S. S. Unpublished notes on the derivations of the confidence limits of the regression intercept.

Manuscript received 2/15/51
Revised manuscript received 6/1/51

PSYCHOMETRIKA—VOL. 17, NO. 1
MARCH, 1952

A FACTOR ANALYSIS OF WOMEN'S MEASUREMENTS TAKEN FOR GARMENT AND PATTERN CONSTRUCTION*

HELEN HEATH
CHICAGO, ILLINOIS

In order to facilitate garment and pattern construction, a research which involved taking a minimum of fifty-five measurements on several thousand women was conducted by the Bureau of Home Economics. Partial correlations with age held constant were computed for a representative group of 4,128 of the women. The correlations among twenty-nine of these variables served as the basis of the present study. By a combination of the multiple-group and the centroid method of factoring, five factors were extracted. After twenty-nine rotations, simple structure was evident, and the factors were interpreted as bone length, size of joints, circumference below the waist, circumference of extremities, and circumference above the waist. The intercorrelations of the primaries were computed, and two second-order factors were extracted. One of these seems to be primarily related to the growth of fatty tissue and the other to the development of the bones.

Prior to the era of psychology there was at least a half-hearted conviction that personality was related through some obscure channels to body build. Motivated by the hope of verifying or refuting this view, hypotheses have been formulated and carefully investigated by members of the Italian School of Clinical Anthropology (6, 13, 16, 25), by Kretschmer (9) in Germany, Sheldon (20, 21, 22) in America, and numerous other students of human behavior. An extended historical review of these studies is presented by Cabot (3).

Within fairly recent times the technique of factor analysis has been applied to this problem. The original plan for the factor analysis studies was to determine whether or not a single factor is sufficient to account for human growth. Later investigations which utilized the multiple-factor methods were designed with the immediate aim of revealing the parameters of physical growth and with the ulterior hope that the causative factors responsible for growth may later be linked with particular characteristics of temperment. Spearman (23), Burt (1, 2), Cohen (4, 5), Mullen(12), McCloy (10), Holzinger and Harman (8), Hammond (7), Rees (17), Rees and Eysenck (18), Thurstone (24, 25) and Moore and Hsü (11) are among the

*The author wishes to express her appreciation to Professor L. L. Thurstone and to the members of the Psychometric Laboratory Staff for advice relative to this study.

authors who have reported studies on factor analysis of anthropometric measurements. Subjects on whom the measurements were taken include normal children, neurotic or delinquent children, adolescents, college students, average male adults, neurotics, psychotics, and criminals. The methods of factoring vary with the preferences of the authors and range from Spearman's general factor method, through the bi-factor methods, multiple-factor methods without rotations, with orthogonal rotations, and with oblique rotations. Furthermore, there is little consistency regarding the nature and number of variables selected. Some studies include face and head measures, some hand measures; and others are predominantly measures of the trunk and limbs.

With such a number of variations, precise uniformity of results could scarcely be anticipated. However, certain factors appear in several of the studies. These will be considered in relation to the factorial composition of the present research which is based on a correlation matrix of twenty-nine measurements taken on 4,128 normal adult women.

Source of the Data

The data for this investigation are contained in a government bulletin (14) which reports a study conducted between July 14, 1939, and June 1, 1940, by the Bureau of Home Economics subsidized by a Federal Project Grant of the Work Projects Administration. The aim of the research was to standardize sizes of women's clothing in order to facilitate garment and pattern construction. A minimum of fifty-five measures were taken on over 14,000 women with normal physique who were residing or visiting in the District of Columbia, Arkansas, California, Illinois, Maryland, New Jersey, North Carolina, or Pennsylvania. Both native and foreign-born women were represented; all, however, were of the Caucasian race. Although members of a variety of economic backgrounds participated, it was the final opinion of those in charge of the measuring program that the average woman measured belonged to a lower-than-average income group. The age of the subjects ranged from eighteen to eighty.

Ten schools were established which provided short courses for training the measurers, and standards of precision were required before one was considered qualified to participate in the measuring. The entire program was carefully supervised and one may have confidence in the accuracy of the results. A detailed statistical study was made on 4,128 cases. Twenty horizontal measures were correlated and factored by the method of principle components. This was the appropriate procedure considering the purpose for which the study was designed, as it indicated which were the most essential girth measures to be considered in garment construction. The primary purpose of the present study was to investigate the underlying domain of growth, and for that reason a factoring and rotational procedure which resulted in simple structure was used. Complete and partial correlations with age held constant had been computed for the group of 4,128 women. Since age influences size even during adulthood, the partial correlations were considered more suitable for the factor analysis. The twenty-nine variables which were selected from the total available list are presented in Table 1.

*Procedure and Results**

The first step in any factor analysis is to select a method of estimating the communalities. For the present study the following formula was used; (26, pp. 300 and 318)

$$h_1^2 = \frac{(\sum r_1 + t_1)^2}{\sum r + \sum t}.$$

The application of this formula requires that a matrix be constructed which consists of the correlation coefficients of three or more variables which correlate highest with the measure, denoted as variable 1, for which the communality is being estimated.

h_1^2 equals the estimated communality for variable 1,
$\sum r_1$ is the sum of the known coefficients in column 1,
t_1 is the highest coefficient in column 1,
$\sum r$ is the sum of the known coefficients in all the columns, and
$\sum t$ is the sum of the highest known coefficients in each of the columns.

The multiple-group method (26, pp. 170-175) was selected for the factoring. Three factors were originally extracted by this technique; the residuals were computed, and a fourth factor was taken out by the grouping method. New communality estimates were made by squaring the entries in each row of the factor matrix, and the four

*Only those tables considered most essential are contained in this article. A microfilmed reproduction of the complete thesis which includes a more extensive set of tables may be procured by ordering Dissertation No. T 883 from the Library Dept. of Photographic Reproduction, University of Chicago, for a fee of $1.30.

factors were then extracted simultaneously by the multiple-group procedure. Residuals were still rather large, but no particular pattern was evident, so the centroid method was used for removing the fifth factor. The highest r in each column of the preceding residual table was accepted for the diagonal entry in the corresponding column in this final operation. Although one residual remained with a value of —.16, all the others ranged between —.06 and .14 inclusive; the standard deviation for the distribution of residuals was .0239. Twenty-nine radial rotations were taken before the planes seemed satisfactorily located. The centroid matrix, transformation matrix, and oblique factor matrix are contained in Tables 2, 3, and 4 respectively. The correlations among the primaries were computed and are recorded in Table 5. Diagonal estimates were made by the procedure described previously, and the matrix was factored by the centroid method. Two factors were extracted and four cycles of factoring were required before the communalities were satisfactorily stabilized. Table 6 contains this second-order centroid. One rotation was made; the transformation and the oblique structure which resulted are presented in Tables 7 and 8. The correlation between the two second order factors is .33

Discussion of First- and Second-Order Factors

Factor A has higher loadings than any of the other factors; it is exceptionally well defined, having only one projection between .12 and .40. The five measurements with significant values are stature, sitting height, arm length, tibiale height, and hip height; the factor is obviously related to the length of bones. Sitting height, which is actually the length of the spine, has a much lower loading than the other four. This variable has the highest projection of any on Factor B, thus suggesting that the length of the back bone is partially determined by influences not affecting the long bones. The nature of this parameter will be discussed more fully in consideration of Factor B. The one low value of Factor A which might possibly be considered in the interpretation is .15 for variable 25, anterior chest width. This result may not be entirely due to error variance, as the anterior chest width is partly determined by the length of the ribs, and it is not unreasonable to assume that length of the ribs depends to some extent upon the same growth processes which are responsible for the other long bones. Excluding Spearman (23) and McCloy (10), all the other authors reported a factor which was primarily represented by vertical measures. In most cases measures other than those depending upon bone length were omitted from the factor; Hammond (7), however found bone length related once positively and once negatively to head measures. Holzinger (8), Rees (17) Rees and Eysenck (18), and Cohen (4, 5) found vertical measures in antithesis to girth measures. Since abnormalities of growth are in many cases dependent upon malfunctioning of the pituitary gland, it is possible that minor variations among normals may have the same causation. It was inevitable that the bone length factor was omitted from Spearman's (23) results, as he was interested only in determining whether or not there was a general growth factor analogous to the "*g*" studied in connection with intelligence. McCloy (10) states that he was tempted to call his general growth factor a linear growth factor, for the linear measures had the highest loadings. Several factor analyses involving individuals at different age levels are reported by McCloy, and he noted that as the age increased, the tendency for the linear measurements to predominate became more apparent. However, all the variables excluding measures of fatty tissue had significant loadings on the factor, so general growth was accepted as the more appropriate designation.

Factor B, which was the most difficult of the five factors to interpret, has only one value which is above .40, sitting height. The next highest, ankle girth, is .34. Stature, wrist girth, and minimum calf girth are in the twenties. Elbow and forearm girth with values of .19 each and armscye at .18 have been included in the interpretation. The low values of the elbow and forearm girth may be due in part to the fact that they also have appreciably high loadings on both D and E. All circumference measures were taken with a steel tape, and although great caution was exercised to insure accurate readings, there was probably a larger relative error on the small girth measures than on the large girth and long bone measures. If so, this increased relative error may further explain the rather low projections. Factor B has been interpreted as cancellous bone size. The composition of the vertebrae and the ends of the long bones which include the protusions at the joints is unlike that of the shafts of the long bones (19). The former, which is known as the cancellous portion of the bone, has an open and spongy structure in contrast to the dense shafts of the narrow-filled bones. The three significant projections on B which are not joints are stature, forearm girth, and minimum leg girth. Since stature is partially determined by the length of the spine, its inclusion in Factor B is not unexpected. Both the minimum calf girth and the forearm girth are measures taken rather close to a joint, and consequently the enlargement of the bone as conditioned by the joint could influence these variables. Although it has a low loading, the armscye girth is included; this is consistent with the interpretation, as the circumference of the armscye is controlled in part by the humerus at the shoulder joint. Two variables which one might expect to have high loadings on Factor B but which did not, are knee girth at tibiale and bent knee girth. These omissions may be explained by the fact that women tend to have fat deposits around the knee which in many cases may overshadow the influence exerted by the bone circumference. A factor similar to this one has not been explicitly mentioned by any of the other authors, and there is little evidence that it is apparent in any of their results.

Factor C will be omitted for the present and discussed later in connection with Factor E, so that the similarities between the two may be clearly pointed out. Factor D contains the girth measures of the extremities. With two exceptions, all the accepted variables have loadings above .30. These two are elbow girth and forearm girth, each with a value of .19. Since both of these variables are also included in Factors B and E, the low projections on D are not too disappointing. The next highest value on D is only .10 for midway thigh girth; consequently the dividing line between the significant and the insignificant loadings is very distinct. Factor D begins with the knee and elbow and includes all girth measures below. This factor has not appeared in any of the other studies, although in many cases the required measures were taken. Thurstone's (24) analysis of Hammond's data gave rise to an extremity size factor; however, the extremities were the head and the hands rather than the forearm and lower part of the leg.

Factors C and E are complimentary in that Factor C includes the girth measurements of the lower half of the body, and Factor E, those of the upper half. Minimum calf girth and ankle girth are omitted from the C factor. This may be due to the fact that these are primarily determined by bone circumference while the other below-the-waist girth measures depend chiefly on fat deposits. Variable 21, which is upper arm girth, has a value of .22 on Factor C. This is nine points lower than any of the other reasonably high loadings, but is high enough to indicate that there is some connection between the fat on the upper arm and that on the lower portion of the body. On Factor E there are no values between .08 and .26 so the demarcation is very clear. Unlike C which excluded ankle girth and minimum calf girth, Factor E includes wrist circumference. In fact every girth measure including and above the abdominal-extension girth has a significant loading on E. The projections of the abdominal-extension girth on these two factors are especially interesting. It has a loading of .35 on Factor C and of .31 on Factor E, thus indicating that it is the dividing line between the upper and lower portion of the trunk. Waist circumference has a slight loading of .17 on C; however, most of its variance is accounted for by E. Weight is about equally represented by the two factors with a loading of .33 on C and .28 on E.

With the exception of Spearman's three analyses (23) and the second study of Hammond (7) all of the interpretations include some kind of girth factor. Some of these characterize the type of individual such as "stocky;" others are opposite vertical measures in bi-polar factors; others are described by adjectives such as fat, cross-sectional, or transverse. However, none of these gives any indication of a division such as that found between C and E; no sharp distinction has been noted between the upper and lower girth measures. A verification of this finding is, nevertheless, apparent in Sheldon's observation (22, p. 809) that growth above and below the waist is often not uniform, thus causing dysplastic individuals. As may be expected, the correlation between Factors C and E is very high.

In order to avoid confusion between the first- and second-order factors, the second-order variables will be identified as A, B, C, D, and E, and the new planes will be denoted as X and Y. Plane X has three high loadings, variables C, D, and E which range in values from .57 to .85. The X factor is obviously interpretedas the fat accumulations on the trunk, arms, and limbs. Had face measures been included in this study, it is possible that plumpness of the face would also have been included in this second-order factor. Variable C, which characterizes the girth measures below the waist, the portion of the body generally containing the largest portion of fatty tissue, has the highest value. This second-order factor is probably the result of the interaction between biological and social determinants. C, D, and E have low positive or low negative loadings on Factor Y.

The remaining two first-order variables, A and B, have projections of .46 and .38 respectively on the Y axis. Y is interpreted as the bone-size factor since it is determined by the length of the bone shafts and by the largeness of the cancellous portion of the bones. According to endocrinologists (15), giantism and dwarfism are due to malfunctioning of the anterior pituitary. If the gland is overactive during early life, bones grow abnormally long; the length is not affected, however, if the overactivity begins later. Instead the joints become excessively enlarged; and if hyperactivity is severe, acromegaly results. The variation between the length of the bones and the size of the joints may be interpreted as a result of differences in time of maturation as characterized by closing of the epiphysus in the long bones and decrease in activity of the anterior pituitary. If these always occurred simultaneously, it may be that the variables in the first-order A and B factors would constitute only a single factor — namely, bone size. Both A and B have practically zero loadings on the X factor. The positive correlation between the second-order factors is accepted as an indication of general growth.

REFERENCES

1. Burt, C. The analysis of temperament. *Brit. J. med. Psychol.*, 1938, 17, 158-188.
2. Burt, C. Factor analysis of physical growth. *Nature*, 1943, 152, 75.
3. Cabot, P. S. de Q. The relationship between characteristics of personality and physique in adolescents. *Genet. Psychol. Monogr.*, 1938, 20, 3-120.
4. Cohen, J. Determinants of physique. *J. ment. Sci.*, 1938, 84, 495-512.
5. Cohen, J. Physique, size and proportions. *Brit. J. med. Psychol.*, 1939-41, 18, 323-337.
6. Di Giovanni, A. Clinical commentaries deduced from the morphology of the human body. Translated from the second Italian edition by J. J. Eyre. London and New York, 1919.
7. Hammond, W. H. An application of Burt's multiple general factor analysis of the delineation of physical types. *Man*, 1942, 42, 4-11.
8. Holzinger, K. J., and Harman, H. H. Factor analysis. Chicago: Univ. Chicago Press, 1941.
9. Kretschmer, E. Körperbau und Charakter. Berlin: Springer, 1921. (Translated into English as Physique and Character by W. J. H. Sprott. London: Kegan Paul, Trench, Trubner, 1925.)
10. McCloy, C. H. An analysis for multiple factors of physical growth of different age levels. *Child Develpm.*, 1940, 11, 249-277.
11. Moore, T. V., and Hsü. E. H. Factorial analysis of biological measurements in psychotic patients. *Hum. Biol.*, 1946, 18, 133-157.
12. Mullen, F. Factors in the growth of girls seven to seventeen years of age. Unpublished Ph.D. dissertation, Department of Education, University of Chicago, 1939.
13. Naccarati, S. The morphological aspects of intelligence. *Arch. Psychol.*, 1921, 6, No. 45.
14. O'Brien, Ruth, and Shelton, W. C. Women's measurements for garment and pattern construction. Miscellaneous Publication No. 454. Washington: U. S. Government Printing Office, 1941.
15. Patten, B. M. Human Embryology. Philadelphia: The Blakiston Co., 1946.
16. Pende, N. Constitutional inadequacies. Translated into English by S. Naccarati. Philadelphia: Lea and Febiger, 1928.
17. Rees, L. A factorial study of physical constitution in women. *J. ment. Sci.*, 1950, 46, 619-632.

18. Rees, W. L., and Eysenck, H. J. A factorial study of some morphological and psychological aspects of human constitution. *J. Ment. Sci.*, 1945, 41, 8-21.

19. Rowntree, C. W. Bones, disease and injuries of. *Encycl. Brit.* (14th Ed.), Vol. 3, p. 845.

20. Sheldon, W. H., Stevens, S. S., and Tucker, W. B. The varieties of human physique: an introduction to constitutional psychology. New York and London: Harper and Bros., 1940.

21. Sheldon, W. H., and Stevens, S. S. The varieties of temperament: a psychology of constitutional differences. New York and London: Harper and Bros., 1942.

22. Sheldon, W. H. Varieties of delinquent youth: an introduction to constitutional psychology. New York and London: Harper and Bros., 1949.

23. Spearman, C. The abilities of man. New York: MacMillan, 1927.

24. Thurstone, L. L. Factor analysis and body types. The Psychometric Laboratory, Univ. of Chicago, No. 24, 1945.

25. Thurstone, L. L. Analysis of body measurements. The Psychometric Laboratory, Univ. of Chicago, No. 29, 1946.

26. Thurstone, L. L. Multiple-factor analysis. Chicago: Univ. of Chicago Press, 1947.

27. Viola, G. La Costitizione Individual. Bolonga: L. Cappelli, 1933.

Manuscript received 1/10/51.
Revised manuscript received 3/1/51.

TABLE 1
Reduced Correlation Matrix*

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Weight														
2. Stature	30													
3. Hip height	18	82												
4. Tibiale height	19	74	72											
5. Total posterior arm length	31	76	74	65										
6. Sitting height	33	67	46	43	44									
7. Bust girth	88	08	01	–05	15	15								
8. Waist girth	87	03	03	01	10	10	90							
9. Abdominal-extension girth	90	10	03	06	15	17	85	90						
10. Hip girth	91	18	08	10	20	26	77	78	86					
11. Sitting spread girth	89	16	08	07	18	24	76	78	86	93				
12. Maximum thigh girth	87	14	05	07	16	23	73	72	80	91	90			
13. Midway thigh girth	83	10	03	07	11	20	69	69	75	84	85	91		
14. Bent knee girth	76	27	16	22	26	26	59	60	65	73	73	72	73	
15. Knee girth at tibiale	75	23	17	23	22	26	58	60	64	73	72	73	75	84
16. Maximum calf girth	77	18	07	11	18	23	61	60	64	73	73	75	77	76
17. Minimum leg girth	60	18	09	11	16	21	45	47	48	57	56	56	57	66
18. Ankle girth	53	30	21	23	30	36	40	42	42	48	45	44	45	57
19. Neck base girth	67	23	14	14	22	21	63	62	58	54	53	51	49	46
20. Armscye girth	80	20	11	12	26	24	77	73	73	70	69	69	64	58
21. Upper arm girth	86	03	–01	–01	08	15	84	82	82	80	78	80	76	63
22. Elbow girth	72	21	11	14	25	24	65	63	63	65	62	64	61	59
23. Forearm girth	81	19	08	09	22	26	73	70	71	74	71	74	72	65
24. Wrist girth	61	31	19	22	34	29	51	52	51	52	49	48	48	57
25. Anterior chest width	59	26	17	20	28	23	55	51	50	49	47	46	45	42
26. Highest bust level width	51	14	10	06	18	13	52	51	47	42	42	40	38	35
27. Posterior chest width	62	14	09	09	18	17	64	62	57	53	51	49	46	39
28. Posterior hip arc	76	13	03	07	16	19	63	64	71	86	80	81	73	63
29. Angle of shoulder slope	–05	05	02	02	–01	06	–10	–09	–07	–05	–05	–03	–03	–02
Variable No.:	1	2	3	4	5	6	7	8	9	10	11	12	13	14

*The initial decimal point has been omitted for all entries.

TABLE 1 (Continued)
Reduced Correlation Matrix*

Variable	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1. Weight														
2. Stature														
3. Hip height														
4. Tibiale height														
5. Total posterior arm length														
6. Sitting height														
7. Bust girth														
8. Waist girth														
9. Abdominal-extension girth														
10. Hip girth														
11. Sitting spread girth														
12. Maximum thigh girth														
13. Midway thigh girth														
14. Bent knee girth														
15. Knee girth at tibiale														
16. Maximum calf girth	75													
17. Minimum leg girth	66	72												
18. Ankle girth	54	55	69											
19. Neck base girth	46	47	37	35										
20. Armscye girth	57	57	44	43	59									
21. Upper arm girth	63	64	47	38	60	79								
22. Elbow girth	58	59	50	47	53	67	70							
23. Forearm girth	65	68	56	48	59	73	80	83						
24. Wrist girth	54	53	60	59	46	55	52	58	62					
25. Anterior chest width	41	41	32	31	50	51	52	46	50	42				
26. Highest bust level width	33	34	26	30	44	45	46	38	41	35	45			
27. Posterior chest width	38	43	32	31	51	50	55	47	52	40	23	32		
28. Posterior hip arc	63	64	50	41	45	59	67	54	62	44	40	34	44	
29. Angle of shoulder slope	–02	–03	–02	–01	00	–08	–04	–01	00	–03	–03	–03	–04	–03
Variable No.:	15	16	17	18	19	20	21	22	23	24	25	26	27	28

*The initial decimal point has been omitted for all entries.

TABLE 2
Centroid Factor Matrix F

	I	II	III	IV	V	h^2
1	.43	.59	.66	.06	—.04	.97
2	.91	—.04	—.12	—.15	—.12	.88
3	.86	—.32	.03	—.13	.05	.86
4	.78	—.21	—.03	—.03	—.02	.65
5	.81	—.08	.03	—.16	—.01	.69
6	.60	.24	—.14	—.07	—.26	.51
7	.17	.62	.66	—.14	.12	.88
8	.17	.51	.74	—.08	.15	.87
9	.22	.54	.71	.00	—.07	.85
10	.31	.55	.64	.24	—.23	.92
11	.28	.51	.67	.25	—.19	.89
12	.27	.50	.66	.29	—.20	.88
13	.24	.48	.64	.35	—.11	.83
14	.42	.47	.42	.44	.08	.77
15	.41	.40	.48	.48	.14	.81
16	.34	.48	.46	.46	.09	.78
17	.34	.62	.09	.48	.16	.76
18	.47	.56	—.01	.27	.15	.63
19	.30	.43	.46	—.16	.17	.54
20	.31	.58	.50	—.11	.05	.70
21	.15	.57	.71	—.02	.09	.86
22	.33	.58	.40	—.01	.14	.63
23	.31	.65	.47	.03	.12	.75
24	.45	.52	.20	.09	.21	.57
25	.33	.36	.35	—.11	.08	.38
26	.20	.38	.32	—.14	.10	.32
27	.22	.41	.41	—.12	.06	.40
28	.24	.47	.56	.22	—.25	.70
29	.02	.01	—.09	.02	—.03	.01

TABLE 3
Transformation Matrix

	A	B	C	D	E
I	.85	.14	.03	.17	—.01
II	—.49	.70	—.13	.30	.37
III	.07	—.56	.50	—.33	.21
IV	—.19	—.26	.46	.57	—.76
V	—.01	—.32	—.72	.67	.50

TABLE 4
Oblique Factor Matrix *V*

	A	B	C	D	E
1	.11	.10	.33	.04	.28
2	.81	.25	—.01	.02	.01
3	.92	—.10	—.01	.00	.00
4	.77	—.01	.04	.04	—.07
5	.76	.09	—.02	.01	.08
6	.40	.43	.07	.00	—.02
7	—.08	.09	.11	—.01	.53
8	—.04	—.06	.17	.00	.47
9	—.03	.03	.35	—.08	.31
10	—.01	.08	.54	—.02	.04
11	—.01	.01	.53	—.01	.04
12	—.03	.00	.55	.01	.00
13	—.06	—.05	.51	.10	—.01
14	.07	.01	.31	.38	—.04
15	.09	—.11	.33	.40	—.05
16	.00	—.02	.33	.38	—.03
17	—.10	.25	.08	.59	—.04
18	.07	.34	—.04	.50	.07
19	.10	.07	—.01	.05	.45
20	.03	.18	.10	.03	.43
21	—.10	—.01	.22	.01	.41
22	.02	.19	.03	.19	.37
23	—.03	.19	.09	.19	.38
24	.12	.22	—.06	.36	.26
25	.15	.10	.03	.04	.33
26	.03	.12	—.03	.03	.36
27	.04	.10	.06	.00	.36
28	—.02	.07	.51	—.05	.00
29	.00	.07	—.02	.02	.—05

TABLE 5
Correlations Between Primaries

	T_A	T_B	T_C	T_D	T_E
T_A	1.00	.22	.12	.18	.09
T_B	.22	1.00	.22	.11	.05
T_C	.12	.22	1.00	.60	.73
T_D	.18	.11	.60	1.00	.48
T_E	.09	.05	.73	.48	1.00

TABLE 6
Centroid Factor Matrix *F*

	X	Y
A	.294	—.383
B	.268	—.315
C	.898	.237
D	.647	.110
E	.689	.350

TABLE 7
Transformation Matrix

	X	Y
X	.781	.330
Y	.625	—.944

TABLE 8
Oblique Factors Matrix *V*

	X	Y
A	—.009	.459
B	.012	.386
C	.849	.072
D	.574	.110
E	.757	—.103

PSYCHOMETRIKA—VOL. 17, NO. 1
MARCH, 1952

A TECHNIQUE FOR CRITERION-KEYING AND SELECTING TEST ITEMS*

JOHN W. FRENCH
EDUCATIONAL TESTING SERVICE

For multiple-choice tests where no *a priori* key exists, the initial selection of a key for maximum validity may be made on the basis of the number of persons choosing each alternative and their mean criterion score. The keying formula is derived. Once the initial keying has been done, further precision in keying and item selection may use, in addition, the mean total test score for persons choosing each alternative. Item-selection formulas suggested by Horst and by Gulliksen for maximizing test validity are both in the form of a ratio, an "item-validity index" divided by an "item-reliability index." The formula derived here is shown to be equivalent to the numerators of these formulas. The expression in the denominators uses the total test score. Although a radical appears in the denominator of Horst's formula and not in the denominator of Gulliksen's formula, both of them select the same items in practice.

Multiple-choice tests such as those of Practical Judgment, Data Interpretation, Word Association, and other objective personality tests can be devised in such a way that there is no predetermined correct response to the items. The alternative for each item to be scored as correct may be selected on the basis of expert judgment. One way to obtain appropriate, if not expert, judgment is to administer the test to a group of individuals having known scores on an appropriate criterion. This method calls for keying as "correct" those alternatives that are selected most frequently by those members of the group having high criterion scores.

The method described here for selection of keyed responses to maximize the correlation between total test score and criterion is based on the available item statistics tabulated in the course of the routine IBM machine item analysis used by the Educational Testing Service. For each alternative of each item are tabulated the number of testees choosing it and the total criterion score summed over testees choosing it. Thus for each item there are as many *N's* and mean criterion scores as their are alternatives. Also available is the total criterion score summed over all testees.

The following notation will be used in deriving a formula for selection of item alternatives:

*The author gratefully acknowledges the suggestions and criticisms of Dr. Harold Gulliksen, Research Adviser at the Educational Testing Service.

X_i = the number of items answered correctly by individual i according to a particular key,
x_i = the deviation from the mean of X_i,
y_i = the deviation score of individual i on the criterion,
N = the number of testees,
K = the number of items in the test,
$\overline{X}$ = the mean test score over all testees,
$\overline{Y}$ = the mean criterion score over all testees,
N_j = the number of testees choosing a particular alternative of item j,
$\overline{Y}_j$ = the mean criterion score for testees choosing a particular alternative of item j, and
S_{ij} = the raw score of individual i on item j (i.e., 1 when a particular alternative is chosen, 0 when that alternative is not chosen).

The problem will be solved here by maximizing the validity coefficient,

$$r_{xy} = \frac{\Sigma xy}{N\sigma_x\sigma_y},$$

by maximizing the numerator Σxy. In the process σ_y will be unaffected as, of course, will N. It must, however, be assumed that any change in σ_x will be small in comparison to the change in Σxy.

$$\Sigma xy = \sum_{i=1}^{N} (X_i - \overline{X})\, y_i \tag{1}$$

$$= \sum_{i=1}^{N} (X_i y_i - \overline{X} y_i)\,. \tag{2}$$

The expression in parenthesis may be expressed as a sum over all items in the test.

$$\Sigma xy = \sum_{i=1}^{N} \sum_{j=1}^{K} \left(S_{ij} y_i - \frac{N_j}{N} y_i \right) \tag{3}$$

$$= \sum_{j=1}^{K} \sum_{i=1}^{N} \left(S_{ij} y_i - \frac{N_j}{N} y_i \right). \tag{4}$$

Since $S_{ij} = 0$ when a particular alternative is not chosen, $\sum_{i=1}^{N} S_{ij} y_i$ calls for summation of y_i only over the N_j individuals marking the particular alternative. Thus,

$$\sum_{i=1}^{N} S_{ij} y_i = N_j \bar{Y}_j . \tag{5}$$

The second term within the parenthesis, on the other hand, is summed over the total group, since it does not contain S_{ij}.

$$\sum_{i=1}^{N} \frac{N_j}{N} y_i = N_j \frac{\sum y_i}{N} = N_j \bar{Y} . \tag{6}$$

Substituting (5) and (6) in (4), we may write:

$$\sum xy = \sum_{j=1}^{K} (N_j \bar{Y}_j - N_j \bar{Y}) \tag{7}$$

$$= \sum_{j=1}^{K} \left[N_j (\bar{Y}_j - \bar{Y}) \right] . \tag{8}$$

In order, then, to maximize $\sum xy$ and hence r_{xy}, it is necessary to maximize the expression in brackets for each item of the test. This can be done readily with the values N_j, $\bar{Y}_j$ and $\bar{Y}$ available from routine item-analysis. For each item the alternative having the largest value for $N_j (\bar{Y}_j - \bar{Y})$ is keyed as correct.

Adkins and Toops (2) suggested the use of point-biserial correlations between alternatives and criterion scores for keying. Formula (8) does very much the same thing as the point-biserial correlation, but tends to avoid the selection of alternatives attracting so few subjects as to be less useful to the total test validity than a more popular alternative with a slightly lower biserial correlation.

Formula (8) makes use only of the item-criterion correlations. Before the initial keying, there is nothing else available. As soon as keying has been done, the tests may be scored and a further refinement of the key and selection of items can take item-test correlations into account. A maximizing function for item selection, using item-criterion and item-test correlations has been developed by Horst (6) and by Gulliksen (4, 5). Both of these assume a key to the total test, so that the item-test correlations are known. Horst's development ends up with an expression in the numerator having the same value as the bracket of formula (8). This is the item-criterion factor or validity index. The denominator of his maximizing function is the item-test factor or reliability index. When present notation is used, Horst's expression (6, formula 8) becomes

$$r_{xy}=\frac{\sum_{j=1}^{K} N_j(\overline{Y}_j-\overline{Y})}{\sqrt{\sum_{j=1}^{K} N_j(\overline{X}_j-\overline{X})}}\cdot\frac{1}{\sqrt{N}\,\sigma_y}. \tag{9}$$

When the present notation is used, the denominator or reliability index in Gulliksen's expression for the test validity reads

$$\sum_{j=1}^{K} r_{jx}\,\sigma_j\,, \tag{10}$$

where r_{jx} is the point-biserial item-test correlation and σ_j is the item standard deviation.

The following algebraic manipulation puts Gulliksen's expression into a form such that it can be compared with Horst's expression.

Applying to (10) the formula for point-biserial correlation (1),

$$r_{\text{p-bis}}=\frac{M_p-M_q}{\sigma_t}\sqrt{pq}, \tag{11}$$

where M_p is the mean score on the total test for persons answering the item correctly, M_q the mean for those answering it incorrectly, σ_t is the standard deviation of scores on the total test, p is the proportion of persons answering the item correctly, and q is the proportion answering it incorrectly, and the formula for the item standard deviation,

$$\sigma_j=\frac{\sqrt{N_j(N-N_j)}}{N}, \tag{12}$$

(10) becomes

$$\sum_{j=1}^{K}\frac{\overline{X}_j-\overline{X}_{(N-j)}}{N^2\,\sigma_t}\Big[N_j(N-N_j)\Big]\,. \tag{13}$$

Now, the mean score on the test for the total group is

$$\overline{X}=\frac{N_j\overline{X}_j+(N-N_j)\overline{X}_{(N-j)}}{N}. \tag{14}$$

Therefore,

$$\overline{X}_{(N-j)}=\frac{N\overline{X}-N_j\overline{X}_j}{(N-N_j)}. \tag{15}$$

Substituting (15) into (13) we obtain

$$\sum_{j=1}^{K} \frac{\bar{X}_j - \dfrac{N\bar{X} - N_j\bar{X}_j}{(N-N_j)}}{N^2\,\sigma_t} \left[N_j(N-N_j)\right]. \tag{16}$$

This may be reduced algebraically to

$$\sum_{j=1}^{K} \frac{1}{N\,\sigma_t} \cdot N_j(\bar{X}_j - \bar{X}). \tag{17}$$

The numerator of Gulliksen's formula may be treated in a parallel way so that his expression for the test validity will now read

$$r_{xy} = \frac{\sum_{j=1}^{K} N_j(\bar{Y}_j - \bar{Y})}{\sum_{j=1}^{K} N_j(\bar{X}_j - \bar{X})} \cdot \left[\frac{\sigma_t}{\sigma_y}\right]. \tag{18}$$

The expression in brackets is a constant, and so may be dropped from the maximizing expression, although it is required in computing the validity, r_{xy}. Similarly the $\dfrac{1}{\sqrt{N}\,\sigma_y}$ in (9) may be dropped from Horst's maximizing expression.

It may now be seen that the maximizing expression from Gulliksen's development is the same as that from Horst's development except that a radical appears in the denominator of the latter.

It has been pointed out (3) that Horst's expression is based on the assumption that the correlation of the group of selected items with the group of unselected items is zero. Gulliksen's expression is based on the assumption that the correlation of the group of selected items with the group of unselected items will be unity. Let it suffice to indicate that the true circumstances call for a condition lying between these assumptions, a correlation between .00 and 1.00, so that the two formulas for the denominator may be considered to represent upper and lower bounds for the true denominator.

In applying these formulas, the first step in keying and selecting items must be to use (8) in order to arrive at an initial key. Neither version of the reliability index can be used at first, since there is no test score until after the initial keying is done.

Refinement in the key, including the keying of more than one response to some items and the keying of no response to some items (i.e., item selection), may, then, be made by the application of the maximizing expressions of formula (9) or (18).

The graphical method of item selection suggested in the articles by both Horst and Gulliksen can be employed. In spite of the difference between formulas (9) and (18), the graphical method suggested by Horst (6) as the best practical approximation selects the *same* items as that presented by Gulliksen (4, 5).

In practice the selection process can end here. Actually several further re-applications of the method, as suggested by Horst and by Gulliksen, may result in successive further changes in the group of items selected, since each successive selection depends upon the items already selected. Further applications of the technique will result in successive approximations to a final solution which is stable or which will oscillate slightly.

REFERENCES

1. Adkins, D. C. Construction and analysis of achievement tests. Washington, D. C.: U. S. Government Printing Office. 1947.
2. Adkins, D. C., and Toops, H. A. Simplified formulas for item selection and construction. *Psychometrika,* 1937, 2, 165-171.
3. Green, B. F., Jr. A note on item selection for maximum validity. To be published.
4. Gulliksen, Harold. Item selection to maximize test validity. Proceedings of the 1948 Invitational Conference on Testing Problems—"Validity Norms and the Verbal Factor," Princeton, N. J.: The Educational Testing Service, 1949.
5. Gulliksen, Harold. The Theory of Mental Tests. New York: John Wiley & Sons, Inc., 1950.
6. Horst, A. P. Item selection by means of a maximizing function. *Psychometrika.* 1936, 1, 229-244.

Manuscript received 2/5/51
Revised manuscript received 6/13/51

PSYCHOMETRIKA—VOL. 17, NO. 1
MARCH, 1952

A FACTORIAL STUDY OF TEMPERAMENT*

MELANY E. BAEHR
UNIVERSITY OF CHICAGO

The theory is advanced that personality factors obtained in the first order may often represent combinations of temperament traits that occur in the experimental population. Under these circumstances an investigation of the second order represents a purification process and yields factors which represent the more basic or pervasive characteristics of the original behavior items included in the factorial study. These second-order factors can be obtained directly in the first order by a careful selection of the variables which enter into the analysis. A second-order analysis was undertaken of the nine factors inherent in three of J. P. Guilford's inventories, and four clearly interpretable second-order factors were obtained. Three of these factors were obtained directly in the first order in a new factorial study of twenty-two behavior items. Attention is drawn to the similarities between these factors and traits of temperament postulated by an independent investigator.

Introduction

Multiple-factor analysis found its original application in the investigation of the intellective and cognitive factors of mind. In this domain L. L. Thurstone identified the underlying functional unities or primary factors of mind. These factors were found to be correlated. When these correlations were factored in turn, they yielded second-order factors. In more recent work Thurstone has drawn attention to the similarity between the most conspicuous second-order factor and Spearman's postulated general intellective factor "*g*" (8, p. 403).

Multiple-factor analysis is finding increasing application in other domains, including that of temperament and personality. The present writer has made a factorial study of temperament which emphasizes the significance of second-order factors in this domain. First, a second-order analysis was made of the factors derived from three of J. P. Guilford's inventories (3, 4, 5), and next a new factorial investigation was undertaken to show how the interpretation of second-order factors could aid in determining the functional unities underlying the domain.

In this study a temperament trait is regarded as composed of

*This paper abstracts portions of the writer's Ph.D. dissertation.

those related behavior characterictics which are relatively permanent for the individual. Personality is regarded as the resultant of the interaction of these temperament traits with the environmental conditions to which they are exposed. Such personality attributes as a person's table manners, habits of personal cleanliness, views about religion, his political affiliations, and his social behavior in a given social "set" are largely determined by environmental conditions and may change from time to time in the light of new experiences. This view does not imply a dichotomy of temperament and personality. Temperament is the raw material from which personality is fashioned; and personality is thus the medium through which temperament traits manifest themselves.

The assumptions made in a factorial investigation of temperament are essentially similar to those made in a factorial investigation of the cognitive domain. We assume that the underlying functional unities of the domain can be described by a finite number of linearly independent factors.

The chief difficulty in factorial investigations of temperament arises at the outset when the investigator is assembling his first selection of personality items to be used in the study. He encounters considerably more difficulty in this respect than the investigator of more concrete and well-defined domains, in which people have been schooled to respond to specific stimuli in specific ways. For instance, we have been taught to respond in a specific way when given a series of numbers to add. The investigator in the intellective domain can therefore be reasonably certain that such a test is measuring essentially the same function in all his subjects. In short, in the intellective domain there has been great stress on the standardization and verbalization of response.

In the domain of temperament the situation is quite different. We have not been uniformly schooled to exhibit certain temperament traits in response to certain stimuli. In addition, the temperament trait finds expression in behavior directly, and is largely unstandardized and unverbalized. Yet in the temperament domain the investigator is usually forced to rely on the subject's verbalized responses to items in personality inventories.

In a sense, each item in a personality inventory is a "test" in a battery. It is clear from the foregoing paragraphs that the investigator assembling his first selection of inventory items to cover this more complex and diversified domain will probably be unable to define each so specifically that it is a relatively pure "test" of a particular temperament trait or that the response to the item will be determined by the degree to which a subject possesses a single temperament trait. The response can be expected to be the resultant of a combination of traits in different strengths. Thus the differentiation of responses achieved by a first-order analysis may well represent the different combinations of temperament traits or "temperament ratios" that occur in the experimental population. It is considered that the interpretation of second-order factors in this domain will provide some leverage on the problem of selecting items for subsequent studies which will allow us to circumvent this first differentiation of inventory items which reflect "temperament ratios."

The argument can be represented as follows. Let P, Q, and R represent three behavior patterns which are relatively permanent for the individual. Let us assume that the responses to the items used in the study are in the majority of instances the resultant of two traits acting simultaneously. If there is a group of items for which the response is determined by elements of P and elements of Q (and this is facilitated when behavior pattern P is often associated with behavior pattern Q in the individual) then the simple structure is likely to reveal a first-order factor comprised of these components which we shall call PQ. By the same reasoning we could obtain a first-order factor PR. It is clear that these factors will be correlated. If the correlations between these primary factors are examined factorially, there should be at least one second-order factor on which both PQ and PR have substantial loadings.

The interpretation of the second-order factor is determined by the common elements in the factors on which it has saturations. The interpretation of this particular second-order factor would thus provide a description of the behavior pattern P.

If the majority of the items in a personality inventory were so ill-defined that the responses were the resultant of three (or more) traits acting together, the interpretation of the second-order factor would be more complex but might still lead to the isolation of a relatively stable behavior pattern. For example, if a second-order factor had saturations on two primaries representing traits PQR and PST respectively, its interpretation would be determined by P. If the combinations of traits in the primaries happened to be PQR and PQS, the interpretation of the second-order factor would be determined by PQ.

Certainly there is no assurance that single traits will emerge in a second-order analysis, but it seems more likely that they will so emerge than that any large number of them will appear in a first-order analysis based on the responses to a first selection of inventory items. On the basis of the arguments advanced it might be advantageous, where possible, to investigate the third- and fourth-order factors. However, the original data employed would seldom, if ever, be stable enough to warrant such a step.

The interpretation of the second-order factor is determined by the essential similarities of the inventory items involved. Having in this way achieved a clear concept of the basic nature of the items, we are in a position to examine critically and revise our original selections of items or "tests" in the battery. This revision should allow us to obtain "purer" items, i.e., items for which the response is determined predominantly by a single trait. It will then be possible to circumvent the first systematization and to obtain some of our original second-order factors directly as first-order factors.

Second-Order Analysis of the Guilford-Martin Data

In order to test the feasibility of the procedures outlined above, the first step was a factorial study to determine whether or not second-order factors derived from personality inventories currently in use could be given clear and psychologically meaningful interpretations. For this purpose Guilford's inventory of scores STDCR (3) and the Guilford-Martin inventories of scores GAMIN (4) and O, Ag, and Co (5) were chosen.

Thurstone (9) has shown that the intercorrelations between the thirteen sets of scores obtained from these three inventories can be described by nine linearly independent factors.* Thurstone named these factors as follows: R (Reflective); S (Sociable); E (Emotionally Stable); V (Vigorous); D (Dominant or ascendant in the sense of social leadership); A (Active); I (Impulsive); X_1 (tentatively designated as Confident); X_2 (left without interpretation). The correlations between these primary factors as given by Thurstone are reproduced in Table 1.

The present writer undertook a second-order analysis of this correlation matrix. The first estimate of the communalities was the highest absolute value of the correlation coefficients in each succes-

*In order to determine how many factors were represented in the 13 scores, Thurstone made the factorial analysis with the test reliabilities in the diagonal cells. Thurstone's factors are therefore described in terms of the saturations on the Guilford scores. It should be noted that since the communalities were not placed in the diagonal cells, this procedure does not constitute a second-order analysis.

sive column of the matrix. The communalities were stabilized after three successive factorings by the centroid method. The final orthogonal factor matrix of four factors is given in Table 2. The fourth-factor residuals are shown in Table 3. The orthogonal factor matrix was rotated to simple structure.* The transformation matrix and the resulting oblique factor matrix are given in Tables 4 and 5 respectively. The oblique factors are labeled A, B, C, and D.

Interpretation of the Second-Order Factors

Factor A has high positive saturations (.79) on Thurstone's S (Sociable), (.73) on X_1 (Confident), and (.62) on E (Emotionally Stable). In addition, there is a smaller, negative loading of —.46 on A (Active). Thurstone's factors, in turn, have their highest saturations on the following Guilford-Martin scores:

Factor A

	Thurstone's Factors	Saturations on Guilford-Martin Scores	
79	Sociable	.66	Agreeableness
		.72	Cooperativeness
73	Confident	.35	Freedom from Inferiority Feelings
		.34	Objectivity
62	Emotionally Stable	.50	Emotional Stability
		.50	Freedom from Depression
−46	Active (negative)	.44	General Activity
		.43	Cooperativeness

The emotionally toned responses in this factor are generally adjustive. The negative loading on Active suggests placidity or an absence of high-pressure or high-strung activity. The easy-going and uncomplicated behavior evident here has caused us to designate this factor *Emotionally Stable.*

Factor B has only two high saturations, one of .85 on Thurstone's I (Impulsive) and one of .80 on Thurstone's D (Dominant or social leadership). Thurstone's factors, in turn, have their highest saturations on the following Guilford-Martin scores:

Factor B

Thurstone's Factors	Saturations on Guilford-Martin Scores	
Impulsive	.60	General Activity
	.45	Rhathymia (Carefreeness, etc.)
Dominant	.55	Ascendance
	.42	Social Extraversion

*One alternative rotation was indicated by the structure which introduced minor variations in two of the factors.

The picture is one of impulsive, carefree, and generally outgoing behavior responses, all of which are facilitated by spontaneous reaction to stimuli. We designated this factor *Primary Function,* a term employed by G. Heymans (6) to describe very similar behavior. His conceptual scheme will be dealt with more fully later.

Factor C has high positive saturations of .52 on Thurstone's V (Vigorous), .49 on X_1 (Confident), .47 on E (Emotionally Stable), and .40 on A (Active). Thurstone's factors, in turn, have their highest loadings on the following Guilford-Martin scores:

Factor C

	Thurstone's Factors	Saturations on Guilford-Martin Scores	
52	Vigorous	.74	Masculinity
49	Confident	.35	Freedom from Inferiority Feelings
		.34	Objectivity
47	Emotionally Stable	.50	Emotional Stability
		.50	Freedom from Depression
40	Active	.44	General Activity
		.43	Cooperativeness

The picture is one of vigorous and confident behavior responses free of the restricting influences of emotional instability. This free and vigorous behavior characteristic is remarkably similar to what Heymans calls *Activity* and this factor is so designated.

Factor D has a saturation of .37 on Thurstone's R (Reflective) and a high negative loading of —.55 on Thurstone's E (Emotionally Stable). Thurstone's factors, in turn, have their highest saturations on the following Guilford-Martin scores:

Factor D

Thurstone's Factors	Saturations on Guilford-Martin Scores	
Reflective	—.76	Thinking Extraversion
	—.41	Rhathymia (Carefreeness)
Emotionally Stable (negative)	.50	Emotional Stability
	.50	Freedom from Depression

The behavior pattern is one of thinking introversion, emotional instability, and depression combined with negative Rhathymia. These emotionally toned responses are, in general, nonadjustive and the designation *Emotionally Unstable* is selected for this factor.

It will be remembered that Factor A was designated Emotionally Stable. The appearance of an Emotionally Stable and an Emotionally Unstable factor in a single study suggests that emotionally adjustive and nonadjustive behavior responses are qualitatively different. This will be discussed more fully later.

The Heymans-Wiersma Conceptual Scheme of Temperament Traits

This conceptual scheme of temperament traits is relatively unknown in the United States, but has enjoyed greater popularity in Great Britain and in some of the dominions. It was devised by a Hollander, G. Heymans, who published his work at the beginning of the century. It was elaborated and refined by a number of his followers, including E. Wiersma (10), whose work included an investigation into the relationship between the temperament traits and the development of character and different personality *Gestalten*.

The Heymans-Wiersma scheme utilizes three variables for a typology. These are: (1) Primary-Secondary Function ; (2) Activity; (3) Emotionality. It is postulated that each of these is a continuous variable and that each occurs in every member of the population but in varying degrees of strength. Heymans defines them as follows:

> In general we call someone Emotional on the basis of the frequency and strength of his affective reactions, in proportion to their causes; Active on the basis of frequency and energy of his activities, in proportion to their motives; Primary or Secondary Functioning according to the degree to which cognitive and affective processes 'perseverate' (German: *nachwirken*), in proportion to their importance. (1, p. 316)

These conceptual traits were used by S. Biesheuvel, Chief Psychologist and Director of the Aptitude Tests Section of the South African Air Force during World War II, as part of a test battery for the selection of pilots. During the course of five years' association with this organization the writer was able to observe and study in some detail the results achieved with this scheme of assessment.

Of the three variables, Primary-Secondary Function is probably most easily described in terms of specific behavior characteristics. An individual at the Primary Function end of this behavior continuum is impulsive, lively, and distractible, since he responds readily to new stimuli. In addition to these characteristics, Biesheuvel (2) states that the primary-functioning individual will show oscillations in his rate of work and that work which demands constant concentration will never appeal to him. The primary-functioning individual will show similar variation in mood, though the prevailing mood will be cheerfulness. Biesheuvel continues further:

> This cheerfulness will be unbraked and therefore far more unrestrained, gay and bubbling over than that of the S.F. . . . The P.F. are on the whole mobile and restless, noisy, quick and on the move. . . . The P.F.'s are impulsive because they react to the stimulus of the moment, the desire or impulse of the moment. (2, p. 7)

These characteristics are well represented in Factor B in our second-order analysis, which was accordingly designated Primary Function.

The Heymans-Wiersma concept of Activity would seem to be the expression of general vigor: mental, physical, or both. Biesheuvel (2, p. 11) writes, "Activity further facilitates enthusiasm and optimism, counteracts over-cautiousness, variability, aggressive expression of the emotions and emotional complexity." In our second-order analysis, Factor C has saturations on Thurstone's Vigorous, Active, Confident, and Emotionally Stable factors. These vigorous and confident behavior responses which are unhampered by emotional complexity describe the central concepts of the Heymans-Wiersma Activity variable, and Factor C was accordingly generalized as Activity.

The Heymans-Wiersma concept of Emotionality cannot be direcly related to the second-order factors obtained in this study. Factor A is more descriptive of adjustive emotional responses and has been called Emotionally Stable, while Factor D describes maladjustive emotional responses and has been called Emotionally Unstable. Whether general Emotionality is the trait underlying both factors or whether the maladjustive and adjustive emotional responses are qualitatively different and relatively permanent for the individual, so that they will always appear as different factors, is a matter for further investigation. If this should prove to be the case, they would be more useful for the description of human behavior than the over-all concept of Emotionality.

In their original inventories Guilford and Martin utilized 511 different personality items. Our analysis has allowed us to describe these by four second-order factors, each of which could be given a psychologically meaningful interpretation.

It is of considerable interest that two of these second-order factors (Primary Function and Activity) are very similar to temperament traits included in a conceptual scheme described by Heymans and Wiersma. The two remaining second-order factors may be related to their Emotionality variable. It is suggested that second-order analyses which produce factors which represent the more basic and pervasive characteristics of the original items in an inventory may provide a fruitful means of comparing and unifying the many and varied first-order temperament factors described by different investigators in this field.

A First-Order Factorial Investigation

A new factorial experiment was designed using tests, or in this instance inventory items, which would cover the concepts embodied in the factors described in the second-order analysis of the Guilford-Martin data and, in addition, the concepts embodied in the Heymans-Wiersma general Emotionality factor in order to investigate the following specific questions:

(1) Can the second-order factors be obtained as first-order factors in a new analysis when the "tests" or items are selected with some knowledge of the temperament traits which determine the responses to the items? In other words, can we obtain relatively pure tests of the traits and so circumvent the first differentiation according to combinations of traits or "temperament ratios"?

(2) Can we obtain a general Emotionality factor directly, or will emotional responses again be differentiated in terms of adjustive or nonadjustive behavior characteristics?

The items finally chosen to cover the concepts embodied in the factors from the second-order analysis and the Heymans-Wiersma Emotionality variable are given below. A number of psychologists collaborated in an attempt to exclude items which were ambiguous or were synonyms of others on the list.

_List of Behavior Items_*

1. Agreeable
2. Cheerful
3. Cooperative
4. Decisive
5. Demonstrative
6. Emotionally Stable
7. Energetic
8. Enthusiastic
9. Even-tempered
10. Happy-go-lucky
11. High-strung
12. Impulsive
13. Lively
14. Persevering
15. Prompt Starter
16. Quick Worker
17. Seeks Company
18. Self-confident
19. Socially at Ease
20. Steady Worker
21. Sympathetic
22. Talkative

*In so far as it was compatible with the concepts to be covered, the items represent socially acceptable behavior. A list composed of the antonyms of these words was treated separately and used in another study.

A modified form of the paired comparison technique was used for the assessment of these items. The 22 items were combined in all possible pairs and presented randomly in a single schedule.

For each pair, the rater was asked to underline that item which, in general, was more descriptive of the behavior of the person he was rating. The rater was urged to make a choice of one word in each pair whenever possible, but was permitted to mark both words of a pair when he considered that they were equally descriptive of the person being rated, and to leave both words of a pair unmarked when he was convinced that neither was in any way descriptive of the behavior of the person being assessed. For each person, the score for each item was the number of times it was underlined in the schedule. Completed schedules were obtained for a sample of 200 subjects.*

The product-moment intercorrelation coefficients were calculated for the 22 items. The correlation matrix is given in Table 6. The communalities were stabilized after two successive factorings by the centroid method. The final orthogonal factor matrix had six columns and is given in Table 7. A frequency distribution of the sixth-factor residuals is given in Table 8. The orthogonal factor matrix was rotated to simple structure. The transformation matrix and the resulting oblique factor matrix are given in Tables 9 and 10.

Four of the six rotated factors allow of a clear interpretation. The remaining factors are given only tentative interpretations. The significant saturations for each of the six factors were as follows:

Factor A

Code Number	Item	Saturation
12	Impulsive	—.56
5	Demonstrative	—.42
10	Happy-go-lucky	—.32
20	Steady Worker	+.47
14	Persevering	+.45
15	Prompt Starter	+.29

This bipolar factor is well represented in both directions and could be designated according to either pole. One end of the bipolarity is described by Impulsive, Happy-go-lucky, and Demonstrative, which are all out-going behavior responses, combined with variability and distractability as opposed to steady perseverance. These are the es-

*Use of the paired comparison method of obtaining judgments, even though modified, may restrict the extent to which the results can be generalized. The method was used in order to minimize the more serious distortions which are often introduced by the halo effect when rating scale judgments are used.

sential characteristics of *Primary Function* and the factor is so designated.

Factor B

Code Number	Item	Saturation
2	Cheerful	+.56
9	Even-tempered	+.46
6	Emotionally Stable	+.42
1	Agreeable	+.30
10	Happy-go-lucky	+.23
11	High-strung	—.60
12	Impulsive	—.56
5	Demonstrative	—.41

One end of this bipolar factor is described by placid, stable, and considered behavior responses, combined with a warm feeling tone denoted by cheerfulness and agreeableness. This pleasant, placid, and stable behavior is similar to that described by Factor A in the second-order analysis, and the present factor is therefore similarly designated *Emotionally Stable.*

Factor C

Code Number	Item	Saturation
17	Seeks Company	—.52
22	Talkative	—.32
4	Decisive	+.42
18	Self-confident	+.42
6	Emotionally Stable	+.38
7	Energetic	+.22

The positive loadings for this factor give a picture of vigorous and confident behavior responses which are unhampered by emotional complexity. These are the central characteristics of Factor C in the second-order analysis (which had saturations on Thurstone's Vigorous, Active, Confident, and Emotionally Stable factors) and of the Heymans-Wiersma Activity variable. The present factor has, in addition, a substantial negative loading on Seeks Company and a smaller negative loading on Talkative. Although these behavior characteristics have not previously been included in a description of the Activity factor, they do not seem to be incompatible with the general pattern of behavior which it represents. It seems possible that the constructively active person will not seek out company and will avoid idle talkativeness. In view of these considerations, this factor is generalized as *Activity*. It must be stressed again that the term "Active" as used by Heymans, Wiersma, and Biesheuvel is not synonymous with energetic. It refers to a broader concept of which general energy is but one aspect.

Factor D

Code Number	Item	Saturation
7	Energetic	+.54
8	Enthusiastic	+.53
13	Lively	+.48
9	Even-tempered	—.37
6	Emotionally Stable	—.32

The behavior pattern is one of stimulability combined with moodiness and emotional instability. This hypomanic behavior is the resultant of a combination of some of the elements of Primary Function and Emotional Instability. It is considered that Factor D is an example of a first-order factor which is a combination of elements from different temperament traits. It is designated *Hypomania* (*Primary Function and Emotionally Unstable*).

The structure was not very clear for factors E and F and we have some reservations concerning the interpretation of these factors, especially of Factor E. Factor E has positive loadings of .60 on 3 (Cooperative), .50 on 21 (Sympathetic), and .39 on 1 (Agreeable) with the smallest loading a negative one, —.34 on 10 (Happy-go-lucky). We may think of this factor as portraying amiability.

Factor F has positive loadings of .51 on 19 (Socially at Ease), .43 on 17 (Seeks Company), and .32 on 18 (Self-confident) with negative loadings of —.55 on 16 (Quick Worker), —.32 on 15 (Prompt Starter), and —.28 on 12 (Impulsive). This factor describes the socially comfortable, pleasure-seeking individual who is slow and dawdling as far as physical output and work is concerned. This "social butterfly" type of personality development is quite common, but because of lack of corroborative experimental evidence this factor cannot be designated with assurance.

It will be seen that no factor was obtained in this analysis which was similar to the general Emotionality factor postulated in the Heymans-Wiersma scheme.

The correlations between these primary factors were calculated by the formula developed by Thurstone (7, p. 138) and are given in

Table 11. These correlations give interesting additional information concerning the factors.

A negative association between Primary Function and Activity has been mentioned by both Biesheuvel (1, p. 333f.) and Heymans (6, p. 271) and is indicated in Table 11 by a correlation of —.504. Heymans (6, p. 271) also considered that Primary Function was positively associated with Emotionality. We obtained a correlation of +.464 between the Primary Function factor and the Emotionally Stable factor, which is probably due to the positive affect inherent in the latter. The Hypomania factor represents an association between elements of Primary Function and Emotional Instability. It would seem that Primary Function can be associated either with positive or negative affect which is consistent with Heymans' observations.

As we should expect, the Hypomania (Primary Function and Emotionally Unstable) factor has a high positive correlation with Primary Function and a high negative with Activity. We should expect the Primary Function elements in the Hypomania factor to be positively associated with the affect component of the Emotionally Stable factor. At the same time we should expect the Emotionally Unstable elements in the Hypomania factor to be negatively associated with the Emotionally Stable factor. These conflicting tendencies are represented by the low correlation of +.186 between these factors.

The correlations between the primary factors are consistent with other investigators' observations concerning the associations between these traits. This fact tends to confirm the interpretations made in this study.

A second-order analysis of the primary factors was not attempted for the following reasons. There were only six first-order factors of which four were interpreted with confidence and two were given only tentative interpretations. Under these circumstances it was considered that the number of meaningful variables which would enter into the second-order analysis would be too small to achieve the overdetermination of factors which is required for an analysis to be scientifically convincing.

Summary

The theory was advanced that, when factorial studies of temperament were based on the responses to inventory items which could be expected to be the resultant of a combination of temperament

traits, the determination of second-order factors represented a purification process. Under these circumstances the second-order factors would be more likely to describe the functional unities underlying the domain than the factors obtained directly in the first-order.

It was shown empirically that the nine linearly independent factors inherent in three Guilford-Martin personality inventories could be described by four clearly interpretable second-order factors. Three of these four second-order factors were obtained directly in the first-order in a new factorial investigation based on the responses to selected behavior items. Finally, it was shown that two of the original second-order factors which were obtained again directly in the first-order analysis were very similar to variables employed in a conceptual scheme of temperament traits described by an independent investigator.

REFERENCES

1. Biesheuvel, S. The nature of temperament. *Transactions of the Royal Society of South Africa*, 1935, 23, 311-360.
2. Biesheuvel, S. The diagnosis of temperament. Unpublished guide for testers currently used by National Institute for Personnel Research, Johannesburg, South Africa.
3. Guilford, J. P. An inventory of the factors STDCR. Beverly Hills, California: Sheridan Supply Co., 1940.
4. Guilford, J. P. and Martin, H. G. The Guilford-Martin inventory of factors GAMIN (Abridged edition). Beverly Hills, California: Sheridan Supply Co., 1943.
5. Guilford, J. P. and Martin, H. G. The Guilford-Martin personnel inventory I. Beverly Hills, California: Sheridan Supply Co., 1943.
6. Heymans, G. Gesammelte kleinere Schriften zur Philosophie und Psychologie. Haag: Martinus Nijhoff, 1927.
7. Thurstone, L. L. Multiple-factor analysis. Chicago: Univ. Chicago Press, 1947.
8. Thurstone, L. L. Psychological implications of factor analysis. *Amer. Psychologist*, 1948, 3, 402-408.
9. Thurstone, L. L. The dimensions of temperament. *Psychometrika*, 1951, 16, 11-20.
10. Wiersma, E. D. The formation of character. *Verhandelingen der Koninklijke Nederlandsche Akademie van Wetenschappen* (Tweede Sectie), Amsterdam, 1938, Deel XXXVII, No. 4, 1-48.

Manuscript received 3/12/51
Revised manuscript received 6/1/51

TABLE 1
Correlations between Thurstone's Primary Factors, R_1

	R	S	E	V	D	A	I	X_1	X_2
R	1.00								
S	—.11	1.00							
E	—.23	.52	1.00						
V	.15	—.03	.05	1.00					
D	.07	.01	.04	.03	1.00				
A	.11	—.37	—.18	.32	—.17	1.00			
I	—.01	—.15	—.10	—.11	.71	—.26	1.00		
X_1	.06	.56	.66	.30	.03	—.16	—.19	1.00	
X_2	—.02	—.14	—.12	—.09	—.19	.04	—.22	—.01	1.00

TABLE 2
Orthogonal Factor Matrix F_1

	I	II	III	IV
R	—.18	—.08	.24	.35
S	.59	.40	—.16	.32
E	.65	.59	.12	—.24
V	—.14	.24	.53	.14
D	.47	—.50	.45	.03
A	—.57	.15	.36	—.16
I	.46	—.76	.30	—.07
X_1	.45	.65	.28	.26
X_2	—.18	.06	—.16	—.01

TABLE 3
Fourth-Factor Residuals

	R	S	E	V	D	A	I	X_1
R								
S	—.04							
E	—.01	.00						
V	—.03	—.02	—.03					
D	—.01	—.01	—.01	—.02				
A	—.01	.02	.01	.03	.02			
I	—.04	—.05	—.01	—.02	—.03	—.01		
X_1	.03	—.01	.02	.01	.01	.06	.03	
X_2	—.01	—.08	—.02	—.05	—.01	—.01	—.01	.07

TABLE 4
Transformation Matrix Λ_1

	A	B	C	D
I	.737	.427	.032	.266
II	.460	—.582	.487	—.241
III	—.098	.685	.831	—.098
IV	.485	.099	—.265	.929

TABLE 5
Oblique Factor Matrix V_1

	A	B	C	D
R	—.02	.10	.06	.37
S	.79	—.12	.00	.06
E	.62	.04	.47	—.55
V	.02	.15	.52	.06
D	.09	.80	.14	—.02
A	—.46	—.07	.40	—.07
I	—.07	.85	—.09	—.03
X_1	.73	—.02	.49	—.06
X_2	—.10	—.22	—.11	.04

TABLE 6*
Correlation Matrix R_2

Code																					
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1																					
2	43																				
3	62	18																			
4	–36	–49	–24																		
5	–29	–01	–28	–24																	
6	29	15	34	26	–39																
7	–31	–12	–26	20	–03	–25															
8	–11	19	–15	–22	23	–35	34														
9	55	27	44	–11	–41	59	–38	–34													
10	06	47	–20	–41	21	–13	–14	19	–05												
11	–39	–37	–37	–03	39	–66	16	08	–57	–05											
12	–39	–25	–34	–16	58	–61	14	20	–54	20	62										
13	–22	15	–34	–29	34	–41	41	51	–45	34	20	34									
14	–08	–43	08	51	–45	20	00	–36	10	–49	–07	–37	–46								
15	–29	–38	–11	34	–29	–11	10	–24	–13	–39	12	–03	–24	43							
16	–26	–37	–15	23	–17	–12	18	–19	–21	–27	17	12	–10	24	55						
17	00	20	–13	–40	27	–24	–08	15	–07	27	02	14	20	–37	–30	–37					
18	–08	–06	–14	42	–21	45	01	–15	12	–09	–34	–35	–18	12	–17	–16	–12				
19	11	14	–01	–02	–02	18	–13	–06	04	11	–23	–21	–01	–22	–32	–44	22	43			
20	–01	–30	17	35	–49	28	–02	–42	25	–49	–20	–46	–47	71	53	31	–40	09	–21		
21	35	20	46	–36	–05	–01	–32	08	22	–06	–14	–13	–16	–09	–22	–14	–06	–19	–03	–06	
22	–22	23	–30	–33	38	–40	07	34	–28	18	19	28	42	–45	–19	–13	39	–26	–05	–48	–04

*The decimal points have been omitted in this table.

TABLE 7
Orthogonal Factor Matrix F_2

	I	II	III	IV	V	VI
1	.28	.63	—.23	.19	.14	.01
2	—.21	.68	.04	.36	—.15	—.05
3	.44	.43	—.35	.16	.33	.11
4	.43	—.47	.47	—.30	.04	—.05
5	—.63	—.09	—.19	—.29	.03	—.12
6	.62	.42	.29	—.13	—.07	—.30
7	—.22	—.39	.38	.29	—.18	.04
8	—.53	.04	.15	.32	.27	.08
9	.54	.56	—.10	.04	—.16	—.10
10	—.50	.33	.10	.05	—.24	—.31
11	—.43	—.59	—.29	—.16	.10	.04
12	—.62	—.43	—.32	—.17	.12	—.30
13	—.71	—.07	.24	.31	.10	—.03
14	.68	—.39	.12	—.03	—.04	.25
15	.37	—.60	—.08	.19	—.25	.06
16	.25	—.59	—.14	.24	—.08	—.23
17	—.49	.29	—.09	—.17	—.27	.31
18	.24	.16	.63	—.39	.04	—.09
19	—.07	.40	.29	—.39	.09	.13
20	.75	—.31	.06	.13	—.17	.19
21	.11	.36	—.38	.14	.28	.07
22	—.61	.03	—.10	.07	—.18	.05

TABLE 8
Frequency Distribution of Sixth-Factor Residuals*
($N = 462$)

Residual	Frequency
—.15	2
—.14	0
—.13	0
—.12	2
—.11	2
—.10	0
—.09	4
—.08	4
—.07	20
—.06	16
—.05	30
—.04	38
—.03	58
—.02	50
—.01	50
.00	52
.01	42
.02	50
.03	14
.04	10
.05	6
.06	6
.07	6

*This is the full table of residuals in which each residual is given twice. There are three residuals with an absolute value greater than .10. A reviewer has pointed out that it may have been possible to extract additional factors. This is so. However, a good structure was obtained for the six factors extracted, and it seems unlikely that the extraction of additional factors would have had a significant effect on the interpretation of the factors in this study.

TABLE 9
Transformation Matrix Λ_2

	A	B	C	D	E	F
I	.261	.227	.227	—.237	.137	—.106
II	—.102	.453	—.086	—.158	.238	.331
III	.345	.321	.439	.434	—.397	.250
IV	.338	.577	—.061	.495	—.002	—.520
V	—.201	—.542	.599	.555	.770	.032
VI	.805	—.118	—.621	.420	.418	.738

TABLE 10
Oblique Factor Matrix V_2

	A	B	C	D	E	F
1	—.02	.30	—.02	—.09	.39	.03
2	.00	.56	—.17	.03	—.01	.03
3	.02	.08	.03	—.01	.60	.01
4	.17	—.15	.42	.03	—.23	.03
5	—.42	—.41	—.11	—.09	—.06	.05
6	—.05	.42	.38	—.32	—.11	—.01
7	.21	—.03	.22	.54	—.12	—.12
8	.03	—.03	.03	.53	.12	.01
9	.01	.46	.00	—.37	.08	.00
10	—.32	.23	—.05	—.13	—.34	—.07
11	—.20	—.60	—.13	.06	.01	—.10
12	—.56	—.56	.02	—.06	—.09	—.28
13	—.03	.01	.01	.48	—.15	—.06
14	.45	.00	.07	.02	.03	.03
15	.29	.03	—.09	—.05	—.22	—.32
16	—.02	—.04	.13	—.05	—.21	—.55
17	.06	.00	—.52	—.07	—.04	.43
18	.05	.09	.42	—.02	—.18	.32
19	.00	—.03	.07	—.01	.09	.51
20	.47	.19	.00	—.05	—.04	—.10
21	—.09	—.01	—.05	.01	.50	.00
22	—.09	—.03	—.32	.05	—.15	.05

TABLE 11
Correlations between the Primary Factors R_s

		Primary Function	Emot. Stable	Activity	Hypo-mania		
		A	B	C	D	E	F
Pr. Function	A	1.000					
Emot. Stable	B	.464	1.000				
Activity	C	—.504	—.072	1.000			
Hypomania	D	.667	.186	—.503	1.000		
	E	—.099	.339	.070	—.326	1.000	
	F	.430	.327	.107	.179	—.198	1.000